

---

SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science  
Department of Computer Science  
MASTER THESIS

---



# Beyond Heatmaps: A Visual Concept-Based Explainable Model via Graph Attention Networks

submitted by  
Anar Amirli  
Saarbrücken  
September 2025

---

**Advisor:**

Md Abdul Kadir  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

**Reviewers:**

Prof. Dr. Daniel Sonntag  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

Prof. Dr. Antonio Krüger  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

Saarland University  
Faculty MI – Mathematics and Computer Science  
Department of Computer Science  
Campus - Building E1.1  
66123 Saarbrücken  
Germany

# Declarations

**Option 2: KI-basierte Sprachmodelle werden als Hilfsmittel zugelassen, die Verwendung wird kenntlich gemacht und dokumentiert** *AI-based language models are permitted as tools, but their use must be disclosed and documented*

## Erklärung Statement

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne die Beteiligung dritter Personen verfasst habe, und dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderweitigen fremden Äußerungen entnommen wurden, sind als solche kenntlich gemacht. Insbesondere bestätige ich hiermit, dass ich alle mittels künstlicher Intelligenz betriebenen Software (z. B. ChatGPT) generierten und/oder bearbeiteten Teile der Arbeit kenntlich gemacht und als Hilfsmittel angegeben habe. Ich erkläre mich damit einverstanden, dass die Arbeit mittels eines Plagiatsprogrammes auf die Nutzung einer solchen Software überprüft wird. Mir ist bewusst, dass der Verstoß gegen diese Versicherung zum Nichtbestehen der Prüfung bis hin zum Verlust des Prüfungsanspruchs führen kann.

I hereby declare that I have written this thesis independently and without the involvement of third parties, and that I have used no sources or aids other than those indicated. All passages taken directly or indirectly from publications or other external sources have been identified as such. In particular, I confirm that I have disclosed and documented all parts of the thesis that were generated and/or edited using AI-based software (e.g., ChatGPT), in accordance with the documentation requirements. I agree that the thesis may be checked using plagiarism detection software, including checks for the use of such software. I am aware that any violation of this declaration may result in failing the examination and lead to losing the right to be examined.

Saarbrücken, \_\_\_\_\_  
(Datum Date) (Unterschrift Signature)

## Einverständniserklärung (optional) Declaration of Consent (optional)

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, \_\_\_\_\_  
(Datum Date) (Unterschrift Signature)

**Erklärung**

Ich erkläre hiermit, dass die vorliegende Arbeit mit der elektronischen Version übereinstimmt.

**Statement**

I hereby confirm the congruence of the contents of the printed data and the electronic version of the thesis.

Saarbrücken,-----  
(Datum/Date)

-----  
(Unterschrift / Signature)



## Acknowledgements

I would like to express my sincere gratitude to my advisor, Md Abdul Kadir, for his invaluable guidance and constructive feedback throughout our extensive discussions, as well as for his patience over the course of this lengthy undertaking. His support was instrumental in shaping this research and bringing the thesis to completion.

I also extend my deepest appreciation to Prof. Daniel Sonntag for giving me the opportunity to join his research group. His work on explainable artificial intelligence and interdisciplinary research was a major motivation for me to join the group, and my time there proved to be highly enriching. I would also like to thank him for his comments during the seminar, which were crucial in shaping the direction of this thesis. Furthermore, I thank all the members of his team who supported me throughout this period in any capacity.

I am especially grateful to my dearest friend, Morgane Brette, for her unwavering support, motivation, and positivity throughout the entire process of completing my thesis. I would also like to thank my dear friends Sardar Sardarli, Frederic Neumann, Joshgun Guliyev, Arne Blickle, and my beloved cats, Gaïa and Mouffe, for their companionship. Their kindness, encouragement, and positivity provided the support I needed, especially during times when the workload felt overwhelming.

Lastly, I thank my dear parents, Ziyafat Shamammadova and Bakhtiyar Shamamedov, for their continuous support throughout my education.

## Abstract

Traditional attribution methods, such as heatmaps, highlight the most important areas of an image affecting model decisions. While they indicate where the model is focusing, they fail to explain what within those regions drives the decision. In contrast, concept-based methods offer more human-interpretable explanations by linking model behaviour to high-level visual concepts. Although concept-based explainability methods built upon Vision-Language Models, Concept Bottleneck Models (CBMs), have shown promise in recent years, particularly in medical imaging, they face key limitations: (i) defining clinically meaningful concepts is challenging, (ii) training requires intensive annotation, (iii) concept localisation often relies on heatmaps, and (iv) the extent to which visual models remain faithful to their associated textual concept descriptions can be spurious. Given the strength of deep vision models in capturing subtle visual cues can be critical for model development, we focus exclusively on the visual modality to retain this potential. Yet, existing visually grounded concept-based methods tend to offer only global, class-specific explanations, often neglecting the interactions between concepts. Furthermore, as these methods are typically trained post hoc, they offer limited interpretability during model development. In this thesis, we propose an alternative approach similar to the core idea of CBMs, while grounding them entirely in the visual modality. To achieve this, we formulate concept graphs using a Graph Attention Network (GAT) in an ante-hoc manner. GATs are well-suited for this task due to their ability to capture rich representations while explicitly modelling the relationships between concepts. Moreover, their effectiveness with a shallow architecture (e.g., two layers) makes them inherently more interpretable. To that end, we propose a unified framework that: (a) identifies meaningful visual concepts through non-negative matrix decomposition, (b) constructs optimal, data-driven concept graph representations to establish a link between visual concepts and model outputs. While our models underperform compared to highly optimised task-specific CBMs, they demonstrate consistent generalisation across selected datasets and, in some cases, outperform baseline CBMs. This thesis discusses both the potential and limitations of the proposed framework and aims to encourage further research in this direction. Ultimately, our work seeks to contribute to enhancing interpretability and trust in AI-assisted decision-making, especially in high-stakes domains such as healthcare.

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Explainability . . . . .	3
1.1.1	What is "eXplainable AI"? . . . . .	3
1.2	Current XAI Landscape . . . . .	5
1.2.1	Attribution Methods . . . . .	6
1.2.2	Concept-based Explainability . . . . .	7
1.2.3	Feature Visualization . . . . .	9
1.3	Motivation and Scope of our Research . . . . .	10
<b>2</b>	<b>Related Work</b>	<b>12</b>
2.1	Attribution Methods . . . . .	12
2.2	Concept-based Explanations . . . . .	13
2.2.1	Methods based on CAV . . . . .	13
2.2.2	Concept Bottleneck Models . . . . .	16
2.3	Explainability in Medical Imaging . . . . .	17
2.3.1	Concept-Based Explainability in Medical Imaging . . . . .	17
2.4	Summary and Research Objectives . . . . .	20
<b>3</b>	<b>Technical Background</b>	<b>22</b>
3.1	Deep Learning . . . . .	22
3.1.1	Neural Networks . . . . .	22
3.1.2	Model Training . . . . .	23
3.1.3	Convolutional Neural Networks . . . . .	26
3.1.4	Graph Neural Networks . . . . .	27
3.2	Matrix Decomposition . . . . .	30
3.2.1	Non-negative Matrix Factorization . . . . .	30
3.3	Evaluation Metrics . . . . .	31
3.3.1	Model Performance . . . . .	31
3.3.2	Concept Quality Evaluation . . . . .	31
3.3.3	Faithfulness Evaluation . . . . .	32
<b>4</b>	<b>Methodology</b>	<b>33</b>

4.1	Concept Basis Generation . . . . .	34
4.2	Learning with Concept Graphs . . . . .	35
4.2.1	Concept Graph Representation . . . . .	35
4.2.2	Graph Modelling . . . . .	38
4.2.3	Final Training Objective . . . . .	40
4.3	Explaining Concept Graphs . . . . .	40
4.3.1	Concept-Level Explanation . . . . .	41
4.3.2	Patch-Level Explanation via Concept Propagation . . . . .	42
4.4	Design Considerations . . . . .	42
4.5	Further Evaluation . . . . .	43
<b>5</b>	<b>Experiments and Results</b>	<b>45</b>
5.1	Dataset . . . . .	45
5.1.1	Data Preparation . . . . .	46
5.2	Experiment Setting . . . . .	46
5.2.1	Backbone Models . . . . .	46
5.2.2	Concept Number and Patch Size . . . . .	47
5.3	Quantitative Evaluation . . . . .	50
5.3.1	Evaluation of Concept Generation . . . . .	50
5.3.2	Performance Trade-off . . . . .	51
5.3.3	Benchmark Comparison . . . . .	52
5.3.4	Faithfulness Analysis . . . . .	55
5.4	Qualitative Evaluation . . . . .	57
5.4.1	General-Purpose Dataset . . . . .	58
5.4.2	Skin Lesion Dataset . . . . .	60
5.5	Limitations . . . . .	67
<b>6</b>	<b>Conclusion</b>	<b>70</b>
6.1	Future Work . . . . .	71
	<b>Bibliography</b>	<b>72</b>
<b>A</b>	<b>Disclosure</b>	<b>83</b>

# List of Figures

1.1	Visual comparison of several attribution methods applied to different input images. Each column corresponds to a different attribution technique (e.g., Saliency [97], SmoothGrad [99], Grad-CAM [95], RISE [86], etc.), and each row shows the explanation for a different image. Such methods aim to explain a model's prediction by producing heatmaps in which warmer regions (e.g., red, yellow) indicate input regions with greater influence on the model's output. While all methods serve the goal of post-hoc interpretability, they vary significantly in the type of signals they capture and the granularity of the explanation they provide. <i>Source</i> : [21]. . . . .	7
1.2	Illustration of the CBM framework. Instead of predicting the target label directly from the input, the model initially predicts a set of interpretable, human-defined concepts $c$ (e.g., bone spurs or beak length), which are then used as intermediate representations for the final prediction. This approach facilitates interpretability by exposing concept-level reasoning. Shown here are two example applications: (top) knee osteoarthritis severity grading using radiographs, and (bottom) bird species classification based on visual attributes. <i>Source</i> : [57]. . . . .	9
1.3	Feature visualization by optimizing for different components of a neural network. Each column shows a distinct target: a neuron, a channel, an entire layer, class logits (pre-softmax), and class probabilities (softmax). These visualizations reveal what each part of the network is responsive to, offering insights into the hierarchical and distributed representations learned by deep models. <i>Source</i> : [82]. . . . .	10
2.1	Overview of the Invertible Concept-based Explanation (ICE) framework [121]. The proposed model comprises two main components: a concept extractor and a classifier, which are connected at an intermediate latent space. The concept extractor maps input images $I$ to a middle-layer activation space $\mathbf{A}$ , which is then linearly reduced by a non-negative matrix factorization-based reducer $R$ into a low-dimensional concept space. Here, $\mathbf{A}$ is flattened into a two-dimensional matrix $\mathbf{V}$ . Accordingly, the decomposition $\mathbf{A} \approx \mathbf{U}\mathbf{W}^\top$ is rewritten as $\mathbf{V} = \mathbf{S}\mathbf{P}^\top + \mathbf{U}$ , where $\mathbf{U}$ captures residual information lost during reduction. The classifier uses learned concept weights to produce explanations in terms of concept contributions. The right side shows example explanations for two classes (cat and dog), highlighting the top-3 contributing features along with visualizations of the corresponding concept activations. <i>Source</i> : [121]. . . . .	14

2.2	Pipeline of the Label-Free Concept Bottleneck Model (Label-Free CBM) [81]. The approach first generates a set of candidate textual concepts using a language model (e.g., GPT-3 [9]), which are then filtered. Given a set of input images, both image and text embeddings are computed using vision-language encoders. A concept matrix $\mathbf{P}$ is formed via inner product between image and text embeddings. A concept bottleneck layer is constructed by aligning model activations with $\mathbf{P}$ , and the resulting concept-based features are used for downstream classification through a sparse linear layer. <i>Source:</i> [81]. . . . .	17
2.3	Overview of the concept-based explanation framework by Patrício et al. [83]. The model uses lesion segmentation to preprocess input images and employs a concept encoder to extract concept-specific activation maps from intermediate CNN features. Interpretability is achieved by mapping these activations to predefined medical concepts using global average pooling and a linear classifier, while multiple loss terms ensure visual coherence, semantic consistency, and spatial alignment with expert annotations. <i>Source:</i> [83]. . . . .	19
2.4	Overview of the MICA framework [7] for explainable skin lesion diagnosis. The model aligns visual and textual concepts at multiple semantic levels—image, token, and concept—to ensure global, localized, and categorical interpretability. Concept activation vectors are first extracted using a frozen CNN and then aligned with visual features using a cross-attention mechanism. During inference, the model outputs concept contributions and generates text- and region-based explanations for clinical interpretability. <i>Source:</i> [7]. . . . .	20
4.1	Overview of our proposed visually grounded concept-based framework. <b>i)</b> Input images are divided into random crops to generate image patches, which are passed through a frozen intermediate layer of a vision backbone. From the resulting patch activations, a concept basis is extracted using non-negative matrix factorisation (NMF), providing interpretable visual concepts independent of class labels. <b>ii)</b> In the second stage, patch activations are systematically aggregated into concept nodes based on the concept basis via a patch pooling mechanism, forming a concept graph. This graph is processed by a Graph Attention Network to model inter-concept relationships. The framework is trained in an ante-hoc manner, enabling interpretable, visually grounded predictions. . . . .	33
4.2	Illustration of the patch pooling mechanism. Patch-level activations are aggregated into concept nodes based on their concept composition coefficients $s$ . The process moves from left to right with stride, where the bold red rectangle indicates the current region of focus. . . . .	36
4.3	Concept graph representation example based on the edge weight matrix $\mathbf{E}_w$ and feature vectors $\mathbf{h}_v$ . Nodes represent concept embeddings, and edges encode pairwise cosine similarity. Edges with weak similarity are removed. . . . .	37
4.4	Simple diagram of the explanation framework. Nodes in the final layer act as bottlenecks, and the contribution of each patch is traced back from these nodes to the patch-level activations. . . . .	41

5.1	Example dermoscopic images from the $PH^2$ dataset. The top row shows <i>Nevus</i> samples, while the bottom row displays <i>Melanoma</i> images. . . . .	47
5.2	Shows the AUC scores of our proposed model under varying numbers of patch sizes used for concept generation and formulation. The concept number in this experiment is fixed to 15. . . . .	48
5.3	Shows the AUC scores of our proposed model under varying numbers of concepts used for concept generation and formulation. The patch size in this experiment is fixed at $70 \times 70$ . . . . .	48
5.4	Shows the reconstruction error, sparsity, and stability scores of our proposed model under varying numbers of concepts used for concept generation and formulation. The patch size in this experiment is fixed at $70 \times 70$ . . . . .	49
5.5	Concept-level fidelity analysis of our framework. <b>Left:</b> Insertion curve. <b>Right:</b> Deletion curve. Results are shown for the <i>Derm7pt</i> and $PH^2$ datasets. Shaded regions around the curves indicate standard deviation. . . . .	55
5.6	Patch-level fidelity analysis of our framework. <b>Left:</b> Insertion curve. <b>Right:</b> Deletion curve. Results are shown for the <i>Derm7pt</i> and $PH^2$ datasets. Shaded regions around the curves indicate standard deviation. . . . .	56
5.7	Fidelity analysis for semantic alignment between our Concept-Graph model and a baseline ResNet-50. The left image shows the result of removing the most important regions—identified by our model—from the input of the baseline CNN. The right side shows the effect of adding those regions. The corresponding drop and rise in prediction confidence, respectively, suggest that our model localizes semantically meaningful regions similarly to the baseline CNN. . . . .	57
5.8	Top-5 representative samples per concept for the <i>Ambulance</i> and <i>Recreational Van</i> classes of ImageNet dataset. The dominant class for each concept is determined based on its <i>discriminativity score</i> . ✖ indicates concepts predominantly associated with <i>Ambulance</i> , ✔ indicates concepts predominantly associated with <i>Recreational Van</i> , and gray – indicates concepts evenly distributed across both classes. The average <i>discriminativity</i> value across discriminative concepts is approximately 0.80, meaning about 80% of activations for each discriminative concept belong to its associated class. . . . .	58
5.9	Concept-based explanation for an <i>Ambulance</i> image. <b>Left:</b> Input image with bounding boxes marking the three most important patches for the prediction; each bounding box is colour-coded according to the most active concept within that region. <b>Center:</b> Normalized contribution scores of the top three concepts; Concepts 3, 4, and 6 contribute 0.39, 0.15, and 0.14, respectively. <b>Right:</b> Representative samples for Concepts 3 (red), 4 (purple), and 6 (pink); the order and framing colours correspond exactly to the bars in the centre plot for concept importance. Below the visualization, we report the ground-truth label and the model’s prediction with confidence. . . . .	59

5.10	Top-5 representative samples per concept for the <i>Melanoma</i> and <i>Nevus</i> classes of the <i>PH2</i> dataset. The dominant class for each concept is determined based on its <i>disentanglement score</i> . <i>✗</i> indicates concepts predominantly associated with <i>Melanoma</i> , <i>✓</i> indicates concepts predominantly associated with <i>Nevus</i> , and gray – indicates concepts evenly distributed across both classes. The average <i>discriminativity</i> value across discriminative concepts is approximately 0.75, meaning about 75% of activations for each discriminative concept belong to its associated class. . . . .	61
5.11	Patch-level localization for correctly classified samples from the <i>PH<sup>2</sup></i> and <i>Derm7pt</i> datasets. Each image is overlaid with bounding boxes highlighting the top three most influential patches identified by the model. These regions correspond to semantically meaningful areas of the lesion, demonstrating that the model focuses on clinically relevant visual cues without spatial supervision. . . . .	63
5.12	Concept-based explanation for a correctly classified <i>Nevus</i> image from the <i>PH<sup>2</sup></i> dataset. <b>Left:</b> Input image with bounding boxes indicating the three most influential patches. <b>Center:</b> Contribution scores for the top three concepts. <b>Right:</b> Representative examples for Concepts 4 (purple), 8 (yellow), and 5 (brown). Bounding box colours correspond to the dominant concept activating in each patch. . . . .	64
5.13	Concept-based explanation for a correctly classified <i>Nevus</i> image from the <i>Derm7pt</i> dataset. Bounding boxes indicate top contributing patches, colored by the most influential concept in each region. . . . .	64
5.14	Concept-based explanation for a misclassified <i>Nevus</i> image from the <i>PH<sup>2</sup></i> dataset, predicted as <i>Melanoma</i> . . . . .	65
5.15	Concept-based explanation for a misclassified <i>Melanoma</i> image from the <i>PH<sup>2</sup></i> dataset, predicted as <i>Nevus</i> . . . . .	65
5.16	Correctly predicted <i>Melanoma</i> case from <i>Derm7pt</i> dataset using high-scoring, mixed-relevance concepts. . . . .	65
5.17	Correctly classified <i>Melanoma</i> image from <i>PH<sup>2</sup></i> , supported by well-aligned class-specific concepts. . . . .	66
5.18	Patch-level localization examples where the model focuses on surrounding skin regions or unrelated visual artifacts instead of diagnostically relevant lesion areas. These cases highlight occasional misattribution of importance, potentially due to low-level visual biases. . . . .	67
5.19	Failure case where the model makes a correct benign classification but the dominant concept is not strongly aligned with the target class. Although the highlighted patches focus on the lesion area, the top contributing concept (Concept 9, <i>✗</i> ) is generally associated with malignant artifacts, sharing only superficial visual similarities with the current lesion. This indicates the model’s difficulty in detecting fine-grained, semantically accurate features, even when the overall prediction is correct. . . . .	68



5.20 Failure case where the model correctly predicts benign class, but the explanation reveals a certain degree of semantic inconsistency. While the top-ranked concept (Concept 4, ✓) aligns well with the target class, a significant contribution also comes from Concept 7 (✗), which is typically associated with malignant features. This highlights limitations in concept disentanglement and suggests that the model may be sensitive to subtle, mixed visual cues. . . . .	68
--	----

---

## List of Tables

5.1	Disentanglement score statistics on PH2 dataset for concept counts from 6 to 12. Best score is given in <b>bold</b> , second-best score is <u>underlined</u> . We set the $\lambda = 1.0$ and top-10% in our experiment. . . . .	50
5.2	Concept extraction comparison on <i>ImageNet</i> . Den, R50, and Mob denote DenseNet-201, ResNet-50, and MobileNet-V2. Concepts are extracted from the final activation layer of the networks. Results are obtained from a set of $\sim 2.5$ k images for two semantically similar classes, <i>Ambulance</i> and <i>Recreational Vehicle</i> . . . . .	51
5.3	Concept extraction comparison on the large-scale dermoscopic image dataset of pigmented lesions, <i>HAM10000</i> . Den, R50, and Mob denote DenseNet201, ResNet50, and MobileNetV2. Concepts are extracted from the final activation layer of the networks. Results are obtained from a set of $\sim 10$ k images from all classes of <i>HAM10000</i> . . . . .	51
5.4	Performance comparison between ResNet-50 and our proposed part-based Concept-Graph bottleneck model using three different backbone architectures on two semantically similar classes ( <i>Ambulance</i> and <i>Recreational Vehicle</i> ) of <i>ImageNet</i> . <u>Underline</u> indicates the best result. The performance is reported as $\text{mean}_{\text{std}}$ across multiple runs on test set. . . . .	52
5.5	Performance comparison between ResNet-50 and our proposed part-based Concept-Graph bottleneck model using three different backbone architectures on the <i>HAM10000</i> skin lesion dataset. Models are trained on the <i>Melanoma</i> and <i>Melanocytic Nevi</i> classes. <u>Underline</u> indicates the best result. The performance is reported as $\text{mean}_{\text{std}}$ across multiple runs on test set. . . . .	53
5.6	Quantitative comparison of our method against state-of-the-art supervised concept-based approaches for skin cancer diagnosis. All methods, including ours, are built upon the ResNet-50 architecture for feature extraction to ensure a fair comparison. Results are reported as $\text{mean}_{\text{std}}$ across multiple runs on the test set. The best score is shown in <b>bold</b> , and the second-best is <u>underlined</u> . The benchmark results are reported as stated in the corresponding publications. . . . .	53
5.7	Localisation ratio $R_\alpha$ across different overlap thresholds $\alpha \in \{0.10, 0.25, 0.50\}$ using patch size 70 and stride 0.5. . . . .	57

---

# Chapter 1

## Introduction

No matter how hard we seek the truth, some mysteries are meant to remain unsolved. Strangely enough, there might even be a certain relief in their unresolved nature—a nature that offers comfort in the face of our ignorance. Yet, this reassurance only holds when the truth exists in indifference to our reality. Artificial Intelligence (AI), however, is not one of those mysteries; its position is far from indifferent, and its impact is more tangible than ever. Like any technology with great potential, AI must be properly understood. The central objective of this thesis is to advance this pursuit. Before we delve into the topic of explainability in AI, however, it is essential to briefly revisit the origins of the field and reflect on why we should not remain silent amid its rapid advancements. It is with this curiosity and sense of responsibility that this thesis was formed, simply asking how things might be done differently and more transparently.

A long time has passed since when ancient Greeks first envisioned something resembling what we now call artificial intelligence; the myth of Talos, a bronze automaton created by Hephaestus to guard the island of Crete. Talos was self-moving and performed tasks autonomously, like hurling rocks at approaching ships, demonstrating an early conceptualization of AI as a machine capable of independent action [104]. It took humanity millennia to achieve the level of technological development prerequisite for contemplating ways of making such myth into a reality. The legend of Talos resurfaced in the modern imagination when the idea of creating thinking machines was formally introduced at the Dartmouth Workshop in 1956 [71]. This important event brought about the birth of AI as a formal research field and established the first concrete foundations for building machines with self-intelligence. From its inception, AI has advanced over several stages of evolution, marked by major technological advancements. The development of Machine Learning, followed by Deep Learning, as subfields of statistical learning has played an important role in driving this progress forward. These developments have driven waves of breakthroughs, each pushing the boundaries of what machine can learn and do autonomously, rendering Turing’s Imitation Game an inadequate measure of machine intelligence.

AI underwent a major transformation with the emergence of Deep Learning [59] around a decade ago. Deep Learning, which relies on constructing deep neural network architectures as the name suggests, has driven seminal progress across various fields

by demonstrating an exceptional capacity to effectively extract complex patterns and behaviours from large datasets. This progress, so revolutionary by nature, has been coined the "Industrial Revolution" of the 21st century by many. Like all transformative breakthroughs, the rapid rise of Deep Learning was too fuelled by several other key factors: the exponential growth of available data, substantial improvements in hardware and software system, and critical advancements in research methodologies.

A major turning point for Deep Learning occurred in 2012, when the Computer Vision (CV) field experienced a transformative breakthrough. The winning model in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), based on deep Convolutional Neural Networks (CNNs), demonstrated an exceptional ability in image classification tasks [58]. With this breakthrough, we witnessed a deep learning-based architecture surpass traditional handcrafted-feature methods by autonomously discovering complex and distinctive patterns from raw pixel inputs. This breakthrough signified a major turning point in computer vision, promoting the rapid and widespread uptake of deep learning techniques in image analysis. The impact of deep learning quickly extended beyond computer vision problems, and since then, it has outperformed conventional techniques and inspired waves of innovation across different fields.

Tremendous success of Deep Learning paradigms have sparked a surge in AI-driven innovations spanning domains such as transportation, security, healthcare, and finance. One of the most notable and meaningful contributions of AI has been in the area of medical image analysis. In this domain, AI has achieved high-level accuracy comparable to that of medical professionals in tasks such as classifying chest X-rays [69] and skin lesion images [42]. AI's potential in automated diagnosis can be particularly useful for several reasons. For example, it can process large volumes of images rapidly, enabling faster diagnoses, which is critical in emergency or high-throughput settings. It can also offer a level of consistency that human clinicians may struggle to sustain, especially under fatigue or time constraints. Furthermore, it can assist in detecting subtle patterns or abnormalities that might be overlooked by the human eye, leading to earlier detection of diseases such as cancer [26]. This emphasises AI's potential not only as a diagnostic tool but as a means to improve healthcare access, especially in underserved areas where specialist expertise is limited. However, despite these achievements, the integration of AI into clinical practice has not yet met expectations, contrary to its widespread use in other critical real-life contexts such as in self-driving cars [120, 84]. And the reason for this is simpler than one might think—patients and clinicians are hesitant to rely on AI systems whose decision-making processes are opaque or not easily understood [64].

While AI technologies offers significant benefits, they often are limited in their capacity to make their decision-making processes transparent to human users. This limitation is especially critical in high-stakes domains, such as those in clinical workflow, where the role AI plays is not indifferent. Such applications have little to no margin for error, where a wrong decision could compromise human health and lead to serious consequences in the short term. Evidently, the need for transparency and trust in high-stakes applications has translated into calls for devising strategies capable of exploring the decision-making processes of AI algorithms [84]. In the past decade, these calls subsequently has led to emergence of a distinct research discipline now commonly known as eXplainable Artificial Intelligence or XAI for short. As stated by European Commission's High-Level Expert Group on AI, this field of research aims to develop explainable AI systems, while preserving high learning performance, that enable humans to understand, trust, and effectively manage the integration of intelligent agents into real-world decision-making contexts [18]. The case for transparency in crucial real-world applications is also underpinned by European Union's General Data Protection Regulation (GDPR),

a policy that upholds the right to access information regarding the underlying logic of algorithmic decision-making processes [33]. This mandate reinforces the need for transparency before AI systems can be deployed in sensitive domains such as healthcare. The central focus of this thesis is a healthcare application—specifically, the classification of skin lesions within the broader domain of medical image analysis. Before turning to the core contributions of this work, however, it is important to first clarify what *explainability* entails and how it can be approached. The remainder of this chapter provides essential background for understanding the thesis’s main theme: *explainability in deep learning*. It serves two key purposes: first, to underscore its critical importance in medical AI systems; and second, to offer a structured overview of the major approaches that inform ongoing discussions in the field. To that end, we introduce key concepts and techniques that will recur throughout the thesis. We also briefly address two guiding questions: *what needs to be explained*, and *why it matters*. Lastly, we present a taxonomy of explainability methods, identify where our approach fits within it, and outline the research objectives that drive this work.

## 1.1 Explainability

We start this section with a simple yet fundamental question: *What* needs to be explained, and *why*? Neural networks use a specialized learning algorithm to train on a dataset. Once they’ve been trained, the model makes inferences by leveraging the parameters it has learned to produce predictions. However, the process from input to output involves many layers of complex operations that become nearly impossible to trace, which is why these systems are often referred to as ‘black-boxes’ [22]. Understanding these ‘black-box’ models lies at the heart of many challenges in the field of XAI, including scientific understanding, safety, and ethical accountability. These challenges, in turn, translates into critical objectives that facilitate the need for explainability—such as building trust in model predictions and assisting in meeting regulatory requirements [21].

In essence, *what* must be explained - or brought to light - is the ‘black-box’ nature of these models: either the internal pathway a model takes to arrive at a specific inference, or the type of input data that activates certain internal mechanisms leading to that inference. Humans have long been using AI systems for certain tasks—such as spam filtering, product recommendation, targeted advertisements, and speech recognition—even before the current focus on explainability. These applications demonstrate that functionality often preceded interpretability in AI adoption. The primary reason why explanations were not actively sought in earlier applications indicates that explainability is not necessary for all AI systems [19]. Doshi-Velez and Kim identify two main reasons why AI explainability is not always necessary: when systems pose low risks (like targeted ads) or when they are well-validated and trusted in real-world applications (like postal code sorting or self-driving), as high performance alone often satisfies stakeholders [19]. Conversely, AI systems operating in critical settings that can directly impact human well-being or safety, such as healthcare, legal, and defense, necessitate objectives like accountability [120]. This is *why* certain AI systems must be equipped with explainability.

### 1.1.1 What is "eXplainable AI"?

The rapid and widespread adoption of AI has made it a common part of our daily lives, much like earlier waves of technological innovation. Following the relentless pursuit

of predictive performance since the first major breakthroughs in deep learning, these models have rapidly evolved into highly complex and autonomous systems—often functioning as "black boxes" that lack transparency in how they make decisions. The success of deep learning methods, however, in the absence of transparency, has acted as a catalyst for the formation of this fast-growing research field. Although initial interest in explainability was grounded in technical concerns (i.e., accountability in failure cases) due to the expansion of deep learning into high-stakes applications, it has since been further advanced by non-technical concerns such as ethical, social, and legal considerations.

The term "eXplainable Artificial Intelligence" as we know today was first introduced by Van Lent et al. in 2004 [107] to describe their system's capability to explain the actions of AI-driven agents within simulation games. Due to the recent breakthrough in AI, present XAI research has evolved and is now conducted by a larger community of scientists from various backgrounds and disciplines, such as machine learning, deep learning, robotics, multi-agent systems, human-computer interaction, and cognitive science. As a result of this expansion in the scope of research, there's no commonly accepted definition thus far of XAI. One of the factors that further complicates this is the abundance of similar terminologies that are often used interchangeably to describe only certain aspects of XAI, e.g., explainability, interpretability, transparency, and understandability [3, 120]. Although these terms do have nuanced differences, it is important to mention that they function as subterms rather than directly expressing XAI. Hence, clearly distinguishing these terms is essential for consistent usage and to avoid confusion throughout the remainder of this manuscript.

- **Explainability:** The ability to explain the internal mechanisms a model follows to arrive at a specific output, as well as to explain the individual output, in a human-understandable manner [37]. It tries to explain what leads the model to arrive at its conclusion. For example, an explanation might show which areas of an image led a neural network to classify it as a "cat."
- **Interpretability:** The extent to which the relationship between input features and output predictions can be understood by human [72]. For instance, linear regression is considered highly interpretable because each coefficient has a clear, human-understandable meaning.
- **Transparency:** The extent to which a model is inherently understandable without requiring additional tools or explanations [65]. It addresses the question "*Is the model's structure and logic naturally clear?*" A decision tree, for example, is transparent because its decision path can be followed step by step.
- **Understandability:** The user's ability to comprehend the model's behaviour, potentially aided by explanations or visualizations [74]. It asks whether a person, given their background, can make sense of the model. This concept is subjective and user-centred, depending on domain knowledge and cognitive abilities.

Having established the main terms, we must now ask: what exactly is meant by XAI? The extensive body of research on this topic makes it nearly impossible to provide a single, all-encompassing definition that covers every application and motivation within the field. Attempting such a universal definition would be an overwhelming and impractical endeavour. Instead, as suggested in prior works [120, 21], "eXplainable Artificial Intelligence" (XAI) is better defined and understood in general terms—based on the

specific goals it aims to fulfil rather than on abstract definitions. This definition of XAI is tailored to specific goals, hence allowing readers to focus on the technical aspects most relevant to the original concerns, particularly in high-stakes applications. Among the various objectives discussed in the literature [48, 72, 11, 91, 115, 4], six key goals emerge as central to the ongoing discourse on XAI, as discussed by Fel [21].

- **Building trust in model predictions.** In medical domains, explainability helps clinicians trust AI-assisted diagnostics by highlighting influential regions in medical images.
- **Understanding important aspects of trained models.** For instance, in robotics, explainability could help engineers understand how a robot’s navigation model prioritizes different sensor inputs when making movement decisions in complex environments.
- **Meeting regulatory requirements and aiding certification.** In healthcare, explainability can help verify that AI-based diagnostic tools make decisions based on clinically relevant factors, supporting their approval by regulatory bodies.
- **Uncovering and correcting model biases.** For example, explainability can reveal if a credit scoring model systematically disadvantages certain socioeconomic groups, enabling adjustments to make the system fairer.
- **Identifying and anticipating potential failure cases.** For example, in manufacturing, explainability can show the conditions that lead to equipment failures, enabling preventive maintenance.
- **Debugging models to improve training with human-in-the-loop intervention.** In autonomous systems, explainability helps developers analyse misinterpretations or errors, refine data and model parameters, and enhance performance and safety.

These objectives reflect the main set of goals within the XAI field and emphasise the complexity of the challenges we face in light of the unchecked advances of AI. Having conceptualised XAI, we will now discuss the current landscape of existing XAI methods.

## 1.2 Current XAI Landscape

Although the shift of interest towards XAI research is relatively recent, a substantial body of work has already been conducted in this domain. This progress is largely driven by the motivation stemming from the need to ensure accountability when adapting AI systems in high-stakes applications in real life. In this section, we provide a brief survey of current XAI methods to familiarise the reader with the general methodologies in the field and to clarify how our proposed methods fit within this landscape. These categorisations are based on the general characteristics of XAI methods. Drawing on the recent surveys [118, 2, 120, 43] of explainability techniques, we can initially characterise these methods using the following general criteria:

- **Global vs Local:** Global explainability methods offer a holistic understanding of the inner workings of deep learning models, whereas local explanations provide insight into individual model predictions [43].

- **Ante-hoc vs Post-hoc:** Ante-hoc methods seek to build characteristics of explainability directly into the model itself, whereas post-hoc approaches provide interpretability after a model has been trained—as the name suggests—without altering its internal workings [43].
- **Model-agnostic vs Model-specific:** Model-agnostic explainability methods are designed to provide explanations for all types of models, regardless of their architecture, whereas model-specific methods offer explainability tailored to a particular model [43].

Local explanations focus on clarifying individual predictions, often by attributing importance to specific features or by using concept-based methods, both of which we will discuss in more detail in the following part of this section. Global explanations, in contrast, aim to uncover the broader mechanisms underlying a model’s behaviour, seeking to reveal the general logic that guides its decision-making. These too often rely on concept-based approaches. Ante-hoc techniques are typically global model-specific, as they are often designed to align closely with the architecture and objectives of a particular model. As a result, they may require additional implementation effort and domain expertise to design and integrate effectively. Post-hoc methods, however, are designed to provide interpretability for already trained models without altering their internal structure. These methods typically provide local explanations and are often model-agnostic by design, allowing them to be applied across a broad range of deep learning architectures.

In general, explanations are either local or global, explaining a single prediction or the model as a whole, respectively [43]. There is also a third emerging mode of explainability that does not necessarily fall into either category: *explaining through data* [56]. This approach seeks to understand model behaviour by assessing the impact of training data on the model’s behaviour, using techniques such as influence functions. Specifically, it enables the assessment of how removing a particular data point would affect the model’s parameters or predictions. Although still a relatively new and evolving area within XAI, it provides valuable insights into how specific data points affect model behaviour and robustness. In the following sections, however, we will concentrate on most common explainability methods to familiarise readers with the current landscape of XAI and to help clarify where our proposed approach fits. Furthermore, since our thesis centres on vision models, the discussion will primarily focus on the most recent advances in explainable AI tailored to image-preprocessing tasks.

### 1.2.1 Attribution Methods

In the early stages of XAI research, as AI models were becoming increasingly opaque and complex, much of the focus was directed toward post-hoc explainability to validate and build trust in model predictions. Local explanation methods, which aim to interpret individual predictions, gained particular traction due to their model-agnostic nature and ease of implementation. These approaches were initially driven by the introduction of attribution methods, which have since become a foundational element in the broader landscape of explainable AI [97].

Attribution methods seeks to uncover the reasoning behind a model’s decision and can be applied to wide range of tasks whether it is classifying an image, detecting a specific object, or making a continuous prediction in regression tasks. It is largely due to their simplicity and compatibility with modern deep learning frameworks that



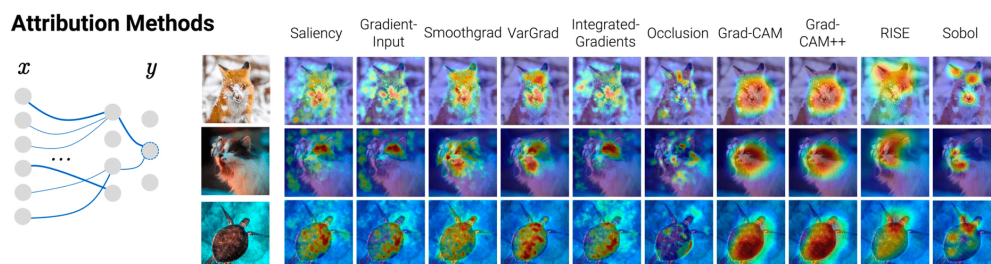


Figure 1.1: Visual comparison of several attribution methods applied to different input images. Each column corresponds to a different attribution technique (e.g., Saliency [97], SmoothGrad [99], Grad-CAM [95], RISE [86], etc.), and each row shows the explanation for a different image. Such methods aim to explain a model’s prediction by producing heatmaps in which warmer regions (e.g., red, yellow) indicate input regions with greater influence on the model’s output. While all methods serve the goal of post-hoc interpretability, they vary significantly in the type of signals they capture and the granularity of the explanation they provide. *Source:* [21].

attribution methods have since become particularly popular. By generating heatmaps that show which input regions contribute most to the prediction, they offer intuitive, visual explanations, making them a widely adopted choice for interpreting model outputs (see Figure 1.1).

Over the years, numerous attribution methods have been developed, utilizing a variety of techniques to generate explanations—ranging from gradient-based approaches [99] and input perturbations [86] to the use of internal model activations [12, 95]. Given a model  $f : \mathcal{X} \rightarrow \mathcal{Y}$  and an input  $\mathbf{x} \in \mathcal{X}$ , an *attribution method* assigns a contribution score to each input variable, thereby explaining the model’s prediction. This functional can mathematically be formulated as:

$$\Phi : f \times \mathcal{X} \rightarrow \mathbb{R}^{|\mathbf{x}|},$$

where  $\Phi(f, \mathbf{x}) = \gamma \in \mathbb{R}^{|\mathbf{x}|}$  denotes the *attribution map* for the model  $f$  for input  $\mathbf{x}$ . Each component of  $\gamma$  quantifies how much each input feature influences the output. A higher value in  $\gamma$  indicates a bigger influence of the corresponding input variable on the model’s decision.

Despite their usefulness, attribution methods face several challenges such as confirmation bias, gradient saturation, and readability, which we will briefly discuss in the beginning of chapter 2.

## 1.2.2 Concept-based Explainability

Another promising and increasingly popular branch of explainability methods which goes beyond traditional attribution methods is *concept-based* explainability. Although these methods initially emerged to address the need for global explainability, they were soon adapted to provide localized explanations as well—making them a powerful technique within the current landscape of XAI methods.

Instead of focusing on pixel-level explanations, these methods emphasize high-level representations grounded in human-interpretable concepts. More specifically, they seek to explain model predictions through abstract concepts that have clear and intuitive

meanings to humans. Among these approaches, *Concept Activation Vectors (CAVs)* [31] and *Concept Bottleneck Models (CBMs)* [57] have emerged as prominent techniques offering highly intuitive and interpretable explanations. Unlike traditional attribution methods, both CAVs and CBMs provide interpretability at both the global and local levels. They not only enable a holistic understanding of the model’s internal mechanisms but also illuminate the relationship between individual inputs and outputs. The primary distinction between CAVs and CBMs lies in their application: CAVs are typically employed as post-hoc explainability tools applied after model has been trained, whereas CBMs are designed and trained in an ante-hoc manner, often requiring domain expertise to define meaningful concepts.

### Concept Activation Vector

Concept Activation Vectors are a post-hoc explainability technique that aims to understand how human-interpretable concepts influence a model’s predictions. Rather than modifying the model architecture or requiring concept supervision during training, CAVs work by analysing existing patterns in the latent activations of a pre-trained model. A concept is determined by defining a vector that represents similar examples within the activation space. This is typically done in one of two ways: (a) applying matrix decomposition, or (b) training a linear classifier to separate those activations from unrelated examples. The resulting vector is the *concept vector* that captures the direction in the activation space that aligns with the concept of interest [52]. This allows researchers to measure how sensitive a model’s prediction is to a given concept, offering insights at both the local and global levels without retraining the model.

Given a model  $f : \mathcal{X} \rightarrow \mathcal{Y}$  and an activation space  $\mathcal{A} \subseteq \mathbb{R}^d$  at a specific layer, a *concept vector* is a vector  $\mathbf{v}_c \in \mathbb{R}^d$  that represents an abstract concept  $c$  within the activation space. The alignment of the input activation  $\mathbf{a} \in \mathcal{A}$  and the concept vector  $\mathbf{v}_c$  shows the degree to which the semantics of any given concept are present in the model’s internal representation of the input.

We will discuss in detail in chapter 2 the existing CAV methods, their usefulness, and how they can be incorporated into concept-based explainability.

### Concept Bottleneck Model

CBMs are a class of interpretable models designed to make predictions based on a set of human-defined, semantically meaningful concepts. Instead of mapping input data directly to the final output, CBMs first predict the presence of predefined concepts and then use these predictions to produce the final output. This two-step structure enables interpretability by design, as the model’s reasoning can be traced through the intermediate concept space. CBMs are typically trained in an *ante-hoc* fashion, meaning interpretability is incorporated from the outset. In particular, the concept space is explicitly defined and embedded as a bottleneck layer during training, requiring annotated concept labels and often domain expertise to construct or validate the concepts.

Formally, CBMs are *ante-hoc* neural networks composed of two stages: a concept encoder  $g : \mathcal{X} \rightarrow \mathcal{C}$ , which maps inputs to a predefined concept space, and a label predictor  $h : \mathcal{C} \rightarrow \mathcal{Y}$ , which produces the final prediction based on the predicted concepts. The concept layer  $\mathcal{C}$  serves as an interpretable bottleneck, allowing users to inspect and potentially intervene in the model’s reasoning process.

CBMs offer a powerful and intuitive approach to interpretability, but they come with

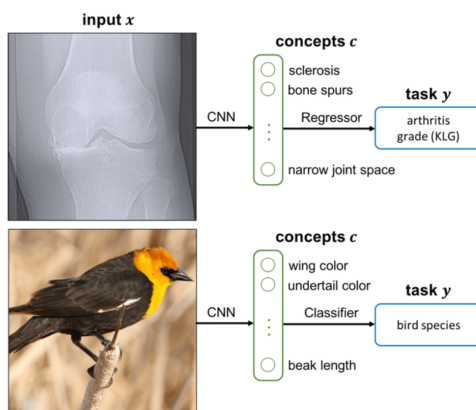


Figure 1.2: Illustration of the CBM framework. Instead of predicting the target label directly from the input, the model initially predicts a set of interpretable, human-defined concepts  $c$  (e.g., bone spurs or beak length), which are then used as intermediate representations for the final prediction. This approach facilitates interpretability by exposing concept-level reasoning. Shown here are two example applications: (top) knee osteoarthritis severity grading using radiographs, and (bottom) bird species classification based on visual attributes. *Source*: [57].

certain limitations. In particular, they require domain expertise to provide accurate concept labels, which can hinder scalability and introduce potential biases. Moreover, they may force models to represent predefined concepts that do not naturally emerge in the learned representation. Despite these challenges, CBMs remain a valuable tool for explainability. We discuss these issues in more detail in chapter 2.

### 1.2.3 Feature Visualization

Feature Visualization is another class of explainability methods that aims to uncover global characteristics of a model by illustrating what individual neurons, layers, or filters in a neural network have learned [78, 82]. It does so by creating images that maximize the activation of selected components within the model. This approach helps reveal the types of patterns the network is sensitive to and provides insight into what each part of the model is detecting in the input data. Visualization methods have been instrumental in understanding feature formation within convolutional neural networks. As post-hoc techniques, they can be easily applied to pre-trained models without requiring changes to the architecture or training process.

Given a model  $f : \mathcal{X} \rightarrow \mathcal{Y}$  and a specific component  $u$  (e.g., a neuron or layer) within the network, *feature visualization* aims to find an input  $\mathbf{x}^* \in \mathcal{X}$  that maximally activates  $u$ . Formally, this can be expressed as the subsequent optimization task:

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathcal{X}} A_u(\mathbf{x}) - \mathcal{R}(\mathbf{x}),$$

here  $A_u(\mathbf{x})$  is the activation of unit  $u$  in response to input  $\mathbf{x}$ ,  $\mathcal{R}(\mathbf{x})$  is a regularization term that supports interpretability.

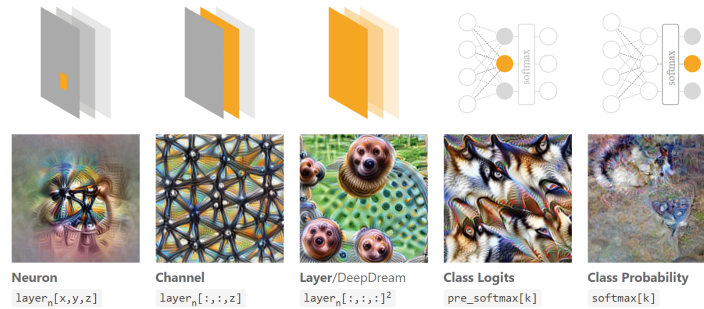


Figure 1.3: Feature visualization by optimizing for different components of a neural network. Each column shows a distinct target: a neuron, a channel, an entire layer, class logits (pre-softmax), and class probabilities (softmax). These visualizations reveal what each part of the network is responsive to, offering insights into the hierarchical and distributed representations learned by deep models. *Source:* [82].

### 1.3 Motivation and Scope of our Research

Despite a substantial body of research, explainability remains an unresolved challenge in the field of medical AI. This is further compounded by the high standards set for interpretability in high-stakes domains like healthcare. Moreover, medical use-cases are highly specific in nature, often involving diverse datasets and requiring domain-specific expertise. Given these concerns, it is vital to develop new strategies that enhance the interpretability of deep learning models before they can be reliably deployed in clinical settings.

One of the most commonly applied explanation methods in clinical settings is attribution methods, owing to their simple, model-agnostic nature. Although they were instrumental in early XAI research, their usefulness in clinical applications has not yet met expectations. This is mainly because such methods typically generate ambiguous heatmaps that lack the clarity required for reliable medical decision-making [49]. The inability of attribution methods to help medical experts understand which specific concepts contribute to a model’s prediction has motivated a shift toward concept-based explainability. These methods offer semantically meaningful and human-aligned insights, making it easier for both clinical experts and patients to understand model behaviour. Additionally, research interest has increasingly focused on developing ante-hoc explainability methods tailored to specific clinical problems.

Although concept-based explainability methods have shown promise in medical imaging, notable challenges remain in effectively embedding clinical knowledge into these approaches [87, 94]. Clinical knowledge is typically incorporated using vision-language models by aligning visual features with textual concept descriptions [83]. Key difficulties involve: (i) the complexity of defining comprehensive and clinically meaningful concepts, (ii) the intensive annotation demands for model training, and (iii) the extent to which visual models remain faithful to their associated textual concepts. Addressing these challenges is vital for advancing the practical deployment of concept-based explainability.

One of the most powerful quality of deep vision models is their ability to learn subtle visual cues that are critical for identifying early pathological patterns—cues that may be overlooked by human experts. Therefore, enforcing strict text-vision alignment, which

prioritises human-understandable explanations, may constrain the model’s ability to detect nuanced features beyond human perception. We aim to retain this capability by developing an unsupervised, concept-based explainability framework grounded entirely in the visual domain. Moreover, recent studies [1, 90] have highlighted the limitations of post-hoc methods in terms of reliability, motivating a shift toward inherently interpretable models. This is particularly relevant in medical applications, where interpretability and trust are essential. In line with this, our approach is designed as an inherently explainable model trained in an ante-hoc manner.

To this end, we construct concept graphs as a form of concept bottleneck model. These graphs are processed using Graph Attention Network (GAT) [110], which capture rich relationships between concepts while preserving interpretability through their shallow architecture and attention mechanisms.

The scope of this thesis encompasses:

- The development of an end-to-end explainable vision model grounded entirely in visual concepts.
- The integration of Non-negative Matrix Factorization (NMF) [61] for unsupervised concept discovery and representation.
- The design of a patch pooling strategy that links local features to concept-level semantics.
- The formulation of concept graphs and their processing using GATs for both classification and interpretability.
- Evaluation on medical image datasets using both standard classification metrics and concept-based explainability metrics.

Our main contributions are as follows:

1. We propose a novel **visually grounded concept-based framework** that does not require textual supervision and offers faithful, structured explanations during training.
2. We introduce a **patch pooling mechanism** that maps image regions to concept activations using a soft assignment strategy based on NMF-derived concept bases.
3. We construct **concept graphs** per image, enabling localized and relational reasoning over visual concepts using GATs.
4. We demonstrate the effectiveness of our framework both on general-purpose and medical imaging datasets, assessing classification performance and interpretability through both quantitative and qualitative analyses.

Ultimately, our work aims to encourage more transparent, interpretable, and trustworthy deep learning systems, particularly in safety-critical domains such as medical diagnosis.

---

## Chapter 2

### Related Work

In this section, we review the existing literature regarding our work on the interpretation of convolutional neural networks using concept-based methods. We explore different interpretation techniques and interaction strategies to illustrate how our research contributes to this topic. However, it is important to note that, since this thesis focuses on concept-based explainability, much of our attention will be dedicated to concept-based explanation methods.

#### 2.1 Attribution Methods

Attribution methods can be thought of as heatmaps that highlight the most relevant parts of image that contribute to the model’s output [97]. These methods generate a spatial importance heatmap by tracking the gradient of the output for a given class to the pixel space. Since the introduction of the first attribution method, Saliency [97], these approaches have been further refined in the context of deep convolutional neural networks for classification studies [12, 99, 102]. However, the gradient of an image only highlights the model’s operation in a very narrow region around that image and can therefore be spurious [29] since the gradient of large vision models is notably large [99]. Other methods, such as Rise [86], Sobol [23] and EVA [24], create attribution maps by probing the output of the model with perturbations in the image to identify areas critical to the decision. Each of these methods employs distinct sampling techniques to investigate the perturbation space surrounding the input in order to achieve this objective. However, attribution methods are subject to other limitations, such as confirmation bias [1, 30, 98] and showing only “*where*” to look at, not “*what*” to look at. These limitations cast doubt on their practical usefulness, as shown by recent studies [17, 38, 53, 80, 96] conducting human-centred experiments to assess the effectiveness of attribution methods.

## 2.2 Concept-based Explanations

Concept-based explanations help users understand model predictions through the lens of concepts that more intuitive for humans to interpret. These concepts can either be direct latent space embeddings associated with high-level features or concept embeddings specified by users, aiming to map the input to an interpretable concept space and use this space to make inferences about the model’s prediction. The former includes associating a set of semantic features with image (CAV-based methods), while the latter involves building a conceptual representation to present semantic features (concept bottleneck models).

### 2.2.1 Methods based on CAV

Part-based explanation methods offers interpretability by directly leveraging high-level semantic embeddings from the model’s latent space. This methods identify what high-level visual features exist and how they contribute to the prediction. Due to representing concepts at high-semantic-level, this methods are inherently more human-understandable and effectively revealing "what" triggers the decision. As an early anchor of concept-based methods, Kim et al. [52] proposed a approach that aims to quantify the influence of hand-picked concepts on the model output. In this work, concepts are defined through so-called Concept Activation Vectors (CAVs), which are an integral part of this method. CAVs, in turn, are obtained by training a simple classifier to distinguish concept-containing examples from unrelated samples within the latent activation space of an intermediate layer. Each CAV’s effect on the model’s output for a given class is then measured to determine the importance of the concepts. To do this, directional derivatives are used to assess how predictions change upon modifications to the inputs in the direction of each concept CAV. Hence, it offers insight into how different high-level concepts impact the model’s outputs. In addition, Kim et al. [52] introduced the first method for evaluating concept importance, called testing with CAVs (TCAV), which measures what fraction of the input images in a given latent space change in the direction of each CAV. Although this method provides a more meaningful explanation to human users than attribution methods, a considerable amount of human effort is needed to construct a database of images representing the relevant concepts [31].

Ghorbani et al. [31] automated concept vector generation with Automatic Concept Extraction (ACE), which extracts CAVs without human supervision. This is done by segmenting input data into meaningful parts with super-pixel segmentation. These segments are first clustered based on their feature representation at an intermediate layer to identify potential concepts. CAVs are then trained for each concept cluster and each cluster is ranked with TCAV to quantify their importance for the model’s prediction. However, one significant drawback of this method is that biases in the explanations may arise from the use of baseline values filled around superpixel segments [39, 45, 100]. In addition, some concept clusters contained background segments, resulting in uninteresting concepts and outliers. To remedy this, an additional cleaning step was introduced to eliminate irrelevant concepts. However, this approach still exhibits certain structural limitations due to its structure. Each image segment can only be assigned to one cluster, users must select an intermediate layer to retrieve relevant concepts, and there is a potential loss of information during the outlier removal phase [25].

Zhang et al. [121] proposed Invertible Concept-based Explanation (ICE), a novel approach to concept-based interpretability. This work applies matrix decomposition meth-

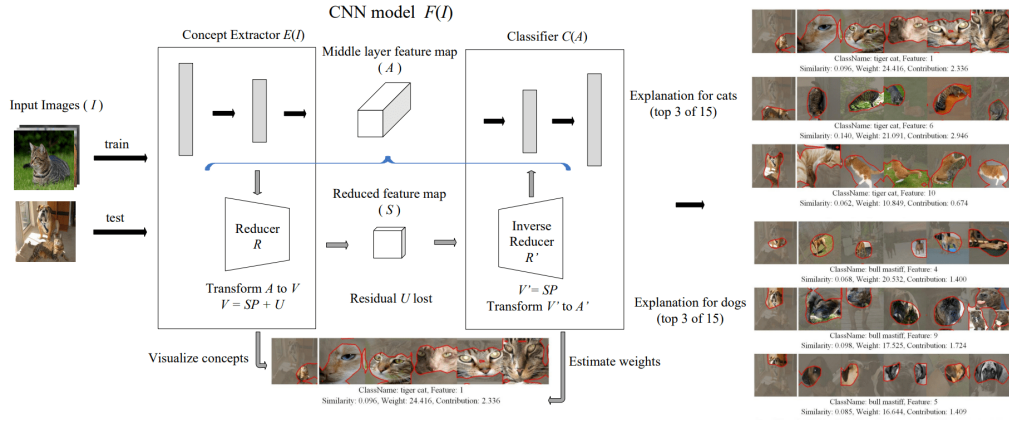


Figure 2.1: Overview of the Invertible Concept-based Explanation (ICE) framework [121]. The proposed model comprises two main components: a concept extractor and a classifier, which are connected at an intermediate latent space. The concept extractor maps input images  $I$  to a middle-layer activation space  $A$ , which is then linearly reduced by a non-negative matrix factorization-based reducer  $R$  into a low-dimensional concept space. Here,  $A$  is flattened into a two-dimensional matrix  $V$ . Accordingly, the decomposition  $A \approx UW^T$  is rewritten as  $V = SP^T + U$ , where  $U$  captures residual information lost during reduction. The classifier uses learned concept weights to produce explanations in terms of concept contributions. The right side shows example explanations for two classes (cat and dog), highlighting the top-3 contributing features along with visualizations of the corresponding concept activations. *Source:* [121].

ods, notably non-negative matrix decomposition method NMF, on internal feature maps to extract CAVs. Here, a feature map  $A$  is extracted as an activation of image segments from an intermediate convolutional layer preceding the global mean pooling. NMF is then used approximate  $A$  with two non-negative low-rank matrices  $W$  and  $U$ ,  $A \approx UW^T$ .  $W$  is the new basis for concept vectors (CAVs) to represent interpretable concepts captured in the feature map, and  $U$  contains coefficients of  $A$  represented with  $W$ . ICE evaluates the importance of the concept using the TCAV score for  $W$ . This approach provides both global explanations and local explanations for a given image. Although the local explanations are directly influenced by the global concept vectors  $W$ , meaning they are not purely local in essence. Another major drawback of ICE is that NFM is directly applied to feature maps at the convolution kernel level, resulting in the localization of concepts. As a result, similar concepts are treated differently in different parts of the image. In addition, it should be noted that other dimensionality reduction methods, including Singular Value Decomposition (SVD) and Principal Component Analysis (PCA), have likewise been investigated as tools for concept discovery [121, 35]. However, by restricting representations to additive linear combinations with non-negative coefficients, NMF yields particularly interpretable concepts, albeit at the cost of lower reconstruction accuracy compared to methods like PCA [22].

While concept-based explanation methods offer meaningful insights to human users, they primarily focus on delivering a global understanding of the model's behaviour. In an effort to move toward more human-interpretable explanations of neural network models on a local level, Y. Ge et al. [28] proposed the Visual Reasoning eXplanation framework (VRX) as an extension to ACE. First, class-specific visual concept clusters are constructed using the ACE method. Image segments are extracted through super-



pixel segmentation and then assigned to a concept cluster as candidates based on the Euclidean distance between their activation and the mean activation of the concept cluster. The segments above a specified threshold are labelled as undetected and filled with a constant value. Once the image segments are assigned to their respective concept clusters, their feature activations are organized into graph convolutional networks (GCNs). In this setup, each node represents a concept, and the edges between nodes capture the pairwise relationships among these concepts. The GCNs are trained to replicate the behaviour of the original model through knowledge distillation. Then the attention mechanism inherent in GCNs is utilized to highlight the significance of visual concepts locally in a given image. While VRX provides intuitive interpretation for human users, it not only suffers from the same severe pitfalls as ACE but is also affected by several other drawbacks. The information is lost due to an additional post-clean-up step that removes uninteresting background regions using an attribution method when creating concept clusters. Another pitfall is that the method for assigning concept candidates to clusters sometimes assigns irrelevant image segments, as it relies on a simple matching based on minimum distances and thresholds. Graph Convolutional Networks (GCNs) exhibit several structural limitations, such as relying on dense graph representations, lacking inter-concept information in edges, and requiring manual intervention to match concepts. Perhaps most critically, GCNs employ a permutation-sensitive approach—even though, in principle, the underlying graph structure should be permutation-invariant [76]. Furthermore, the graph-based representation does not encompass the entire image, since only the closest segment is assigned to each concept centre and fed into the corresponding node. This approach can lead to a loss of information, particularly when the same concept appears multiple times within an image.

Fel et al. [25] introduced the first concept-based explainability method capable of producing concept attribution maps by backpropagating concept coefficients into pixel space, thereby identifying the specific pixels in an input image associated with a given concept. This method known as Concept Recursive Activation FacTORIZATION (CRAFT) is therefore filling the major gap in the attribution methods by indicating both "what" and "where" to look at in images. Once concepts are discovered with NFM, the gradient of concept coefficients  $\mathbf{U}$  with respect to pixels  $\mathbf{X}$  is calculated to generate a spatial heatmap, hence unlocking concept attribution maps. Although CRAFT leverages the same NFM technique used by Zhang et al. [121] on internal feature maps to discover concepts, it is applied after the global average pooling to produce location-invariant concepts. Moreover, they use random crops to extract segments rather than the commonly used super-pixel segmentation technique to provide semantically more coherent segments and circumvent the biases introduced through the baseline value filled around super-pixel segments. Additionally, CRAFT employs Sobol indices on a concept basis  $\mathbf{W}$  to assess the global contribution of concepts to a given object category. Fel et al. [25] shows that solving the concept attributions  $\nabla_{\mathbf{X}} \mathbf{U}$  requires two-step backpropagation steps: lower stage feature extraction  $\mathbf{A} = h_l(\mathbf{X})$  from images  $\mathbf{X}$ ; and the upper stage NMF decomposition  $\mathbf{A} \approx \mathbf{U}\mathbf{W}^\top$ . Hence for the given concept  $i$  the chain rule yields  $\nabla_{\mathbf{X}} \mathbf{U}_i = \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \nabla_{\mathbf{A}} \mathbf{U}_i$ . While the lower stage can be directly computed using backpropagation, the upper stage utilizes implicit function theorem to calculate  $\nabla_{\mathbf{A}} \mathbf{U}_i$  without the need to explicitly backpropagate through each iteration of the NMF solver. As demonstrated by Fel et al. [25], the implicit differentiation of the NMF block  $\nabla_{\mathbf{A}} \mathbf{U}_i$  is integrated into the classic backpropagation  $\frac{\partial \mathbf{A}}{\partial \mathbf{X}}$  to obtain  $\nabla_{\mathbf{X}} \mathbf{U}_i$ . This enables the identification of image regions that activate or contribute to the model's recognition of a specific concept. Since the problem mainly amounts to solving the implicit differentiation of the NMF solver, it is also possible to employ perturbation-based attribution methods [23, 24, 86]

in the lower stage. However, to generate its concept-based attributions, CRAFT relies on the classic attribution method situated on the left-hand side of the chain rule—an approach we discussed above as being vulnerable to a range of issues. While CRAFT represents a significant step towards more human-understandable attribution methods, it is susceptible to similar shortcomings found in attribution methods due to contextual dependence present in different scenarios. Some features might be highly relevant in one instance but not in another due to their relation to other features. This variability can lead to different importance levels for the same feature across different scenarios. Therefore, the absence of contextual information can result in misleading or causally incoherent explanations [112, 117, 73].

### 2.2.2 Concept Bottleneck Models

In recent years, concept bottleneck models have regained attention as an essential line of research for building inherently explainable models. The rationale behind the CBMs is to use human-specified concepts as an intermediate step in determining the final prediction [84]. Introduced by Koh et al. [57], this method first trains neural network to predict intermediate concepts from the intermediate activation of input data. Then, it uses the concept predictions as a new intermediate layer to produce the final output. By directly incorporating human-understandable concepts in the learning process, these models aim to improve the model’s interpretability, as predictions can be traced back through these concepts to understand the model’s reasoning. As mentioned by Koh et al. [57], CBMs are trained in an end-to-end manner, with supervision applied to both the intermediate concepts and the final class predictions. Therefore, despite their advantage in terms of interpretability, CBMs rely on human-provided labels, which limits their scalability and may potentially introduce annotation bias.

To overcome the limitations posed by CBMs, Label-Free CBMs [81] were introduced to transform pre-trained black-box models into CBMs in an unsupervised manner. This method involves automatically creating and filtering a set of concepts using Large Language Models (LLMs) and then employing vision-language text embeddings to align the intermediate activations of a black-box model with the concepts embedding. It enforces an additional sparsity constraint at the last layer from the bottleneck layer to the final output to limit the number of latent concepts involved in the predictions to ensure that only the most relevant concepts contribute to the model’s prediction. This technique aligns with cognitive reasoning process of humans, where reasoning typically relies on only a small set of clearly distinguishable concepts rather than a dense combination of many overlapping factors. While Label-Free CBMs significantly reduce human annotation effort and maintain interpretability through the learned concepts, the interpretability of these automatically extracted concepts may vary, potentially resulting in less semantically coherent concepts. It moreover relies on a critical assumption: the pre-trained model already implicitly encodes the predefined textual concepts. Therefore, the faithfulness of existing concept predictors to their underlying concepts remains challenging. Some studies have shown that correlated concepts may lead to accurate but uninterpretable models that fail to learn localities [87]. This, in turn, makes CBM interpretability fragile, as they occasionally rely upon spurious semantic features.

Schrodi et al. [94] developed unsupervised concept bottleneck models directly from the latent space representations to circumvent the limitations of traditional CBMs regarding their reliance on predefined concepts and the faithfulness of concept predictions. Contrary to conventional CBMs that require priori concept selection, this method relies on

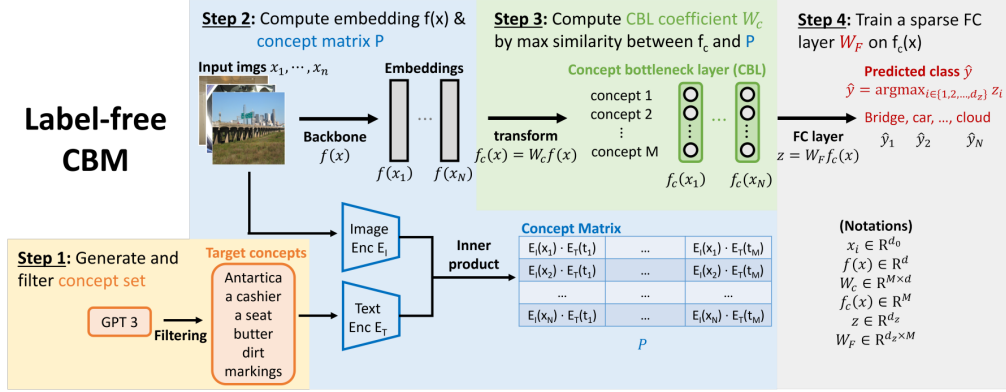


Figure 2.2: Pipeline of the Label-Free Concept Bottleneck Model (Label-Free CBM) [81]. The approach first generates a set of candidate textual concepts using a language model (e.g., GPT-3 [9]), which are then filtered. Given a set of input images, both image and text embeddings are computed using vision-language encoders. A concept matrix  $P$  is formed via inner product between image and text embeddings. A concept bottleneck layer is constructed by aligning model activations with  $P$ , and the resulting concept-based features are used for downstream classification through a sparse linear layer. Source: [81].

NMF for unsupervised concept discovery. Particularly, instead of identifying concepts through vision-language alignment, it applies NMF in the latent space to automatically extract relevant concepts directly from the activations of pre-trained black-box models. Doing so does not force the pre-trained models to encode predefined concepts that they might not inherently represent. In addition, they introduced an additional input-dependent concept selection mechanism combined with sparsity regularization to restrict the prediction to depend on only a few concepts, reducing conceptual complexity among predictions and increasing interpretability. Their experiments show that using direct concept representation with NMF achieves superior downstream performance compared to traditional CBMs relying on concept prediction, while using fewer concepts for prediction.

## 2.3 Explainability in Medical Imaging

### 2.3.1 Concept-Based Explainability in Medical Imaging

The widespread adaptation of large deep-learning models has significantly improved automated medical diagnostics. Despite reaching very high accuracy levels that are often better than the performance of human experts, the clinical deployment of AI-based diagnostic systems in real life still remains limited, predominantly due to a lack of trust regarding the interpretability challenges associated with deep neural networks. This gap of trust has brought about an important research area focused on enhancing transparency, trustworthiness, and interpretability of AI-systems clinical settings. In that line, concept-based explanation methods have become particularly promising for addressing these interpretability challenges. Unlike traditional pixel-based attribution methods—which generate heatmaps of influential pixels but often yield fragmented

and contextually ambiguous interpretations—concept-based methods provide explanations aligned with clinically meaningful visual concepts. These approaches decompose complex model predictions into structured, understandable concepts that resonate with clinicians’ established reasoning processes, helping to foster trust and practical clinical integration.

The part-based framework is one of the earliest adopted approaches for evaluating and utilising concepts in deep learning. Lucieri et al. [66] applied CAVs to skin lesion classification to extract visual concepts, correlating model predictions with dermatologically relevant visual concepts to provide interpretable insights aligned closely with diagnostic reasoning. Graziani et al. [34] employed a similar idea by introducing Regression Concept Vectors. Specifically, they extended the use of CAVs beyond classification tasks to regression settings by introducing Regression Concept Vectors (RCVs). Rather than identifying a discriminative boundary between two concepts, or between a concept and random examples, RCVs aim to capture the direction in feature activation field that aligns with the greatest increase in a continuous concept measure. Hence allowing to understand how features influence predictions and how adjusting certain features might change outcomes. Building on the previous idea, Lucieri et al. [67] proposed a multimodal explanation framework, ExAID, for the explainable diagnosis of skin lesions. This system, too, leverages CAVs to connect latent model activations to clinically meaningful concepts defined by domain experts. Building on these connections, ExAID generates textual explanations that describe which medical concepts were activated during prediction and localise their location on the image using concept localisation maps. This dual-modality allows users to understand what given concepts influenced a prediction and where they are primarily activated within the image. Importantly, ExAID was designed with clinical applicability in mind; it includes a diagnostic interface that presents the explanations in a form usable during real-world workflows.

Patrício et al. [83] advised an inherently interpretable concept-based framework for skin lesion diagnosis. The convolutional kernel is trained to develop concept activations for each concept of interest from the intermediate activations of the backbone model. These concept activations are generated by incorporating an additional loss term to enforce visual coherence in the concept encoder, along with a hard-attention mechanism to align the activations with expert-identified regions relevant to each concept. Concept activation maps are transferred into concept identifiers with global average pooling and are transformed to the output with linear transformation to make the final prediction. Explainability is then attained by calculating the contribution of each concept to the prediction and localizing its presence in the image using concept activation maps. While effective, this method relies on a concept’s annotations for each image by experts and, therefore, is not scalable.

Vision Language Models (VLMs) such as CLIP use a contrastive learning method to learn joint representation from image-text pairs [85]. Although these models exhibit remarkable zero-shot learning abilities across diverse general tasks, their deployment in the biomedical domain remains challenging due to distributional shifts and specialized vocabulary [85]. In recent years, several domain-specific VLMs have been developed to address these challenges in medical research, such as BiomedCLIP [122], PubMedCLIP [20], and MedCLIP [114], which are trained on large-scale medical image-text datasets derived from medical articles [85]. However, despite their broad applicability, these pre-trained models often underperform relative to task-specific models and, furthermore, lack interpretability and transparency in their decision-making processes. [85]. Nevertheless, due to VLMs’ impressive zero-shot learning capabilities in general tasks, CBMs have emerged as a promising type of inherently interpretable model that aims to

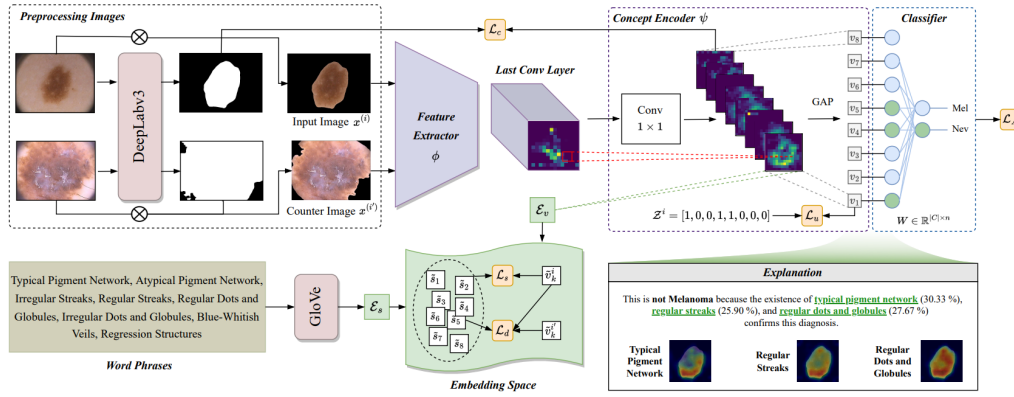


Figure 2.3: Overview of the concept-based explanation framework by Patrício et al. [83]. The model uses lesion segmentation to preprocess input images and employs a concept encoder to extract concept-specific activation maps from intermediate CNN features. Interpretability is achieved by mapping these activations to predefined medical concepts using global average pooling and a linear classifier, while multiple loss terms ensure visual coherence, semantic consistency, and spatial alignment with expert annotations. Source: [83].

base machine learning predictions on human-understandable concepts utilizing VLMs. Rather than relying solely on end-to-end learning from raw input to output, CBMs introduce an intermediate concept layer where the model first predicts the presence of semantically meaningful concepts and then bases its final decision on those concept predictions. This two-stage architecture offers the potential for improved transparency, as the reasoning behind a prediction can be traced through clearly defined, interpretable features.

Several recent studies have employed CBMs to make them more practical for medical settings. These models enable inherent interpretability by basing their final predictions on a predefined set of human-understandable concepts [70, 85]. This capability, however, often entails a substantial annotation effort. To address these challenges, Patrício et al. [85] offers a two-stage methodology that simulates the two stages of a CBM. They employ a pre-trained VLM to automatically predict clinical concepts, then rely on an LLM instead of a sparse layer to make diagnoses based on those predictions.

Wang et al. [113] proposed a CBM-based method tailored for skin disease diagnosis that enhances concept learning by combining predefined concept annotations with the discovery of complementary, data-driven concepts. Unlike most existing CBM approaches that rely on shared image features across all concepts, their method introduces concept-specific adapters that use multi-head cross-attention to detect concept-relevant features individually. This design improves both interpretability and predictive performance by allowing the model to attend to distinct visual cues for each concept.

Bie et al. [7] proposed MICA (Multi-level Image-Concept Alignment), a concept-based framework for explainable skin lesion diagnosis that addresses key limitations of conventional CAV- and ACE-based methods. Unlike prior approaches that typically align visual concepts at a single level, MICA performs alignment across three semantic levels: image-level, token-level, and concept-level representations. At the image level, the model ensures global alignment between entire image embeddings and diagnostic concept embeddings. At the token level, localized visual segments are aligned with relevant

clinical terms, offering fine-grained, region-specific interpretability. At the concept level, MICA enforces semantic consistency between visual evidence and the broader diagnostic categories. This hierarchical structure enables interaction with the model at varying levels of abstraction, supporting both detailed visual cues and high-level textual explanations. While MICA demonstrates strong performance in terms of both interpretability and predictive accuracy across multiple skin lesion datasets, its reliance on manually annotated concept labels for each image limits scalability in real-world applications.

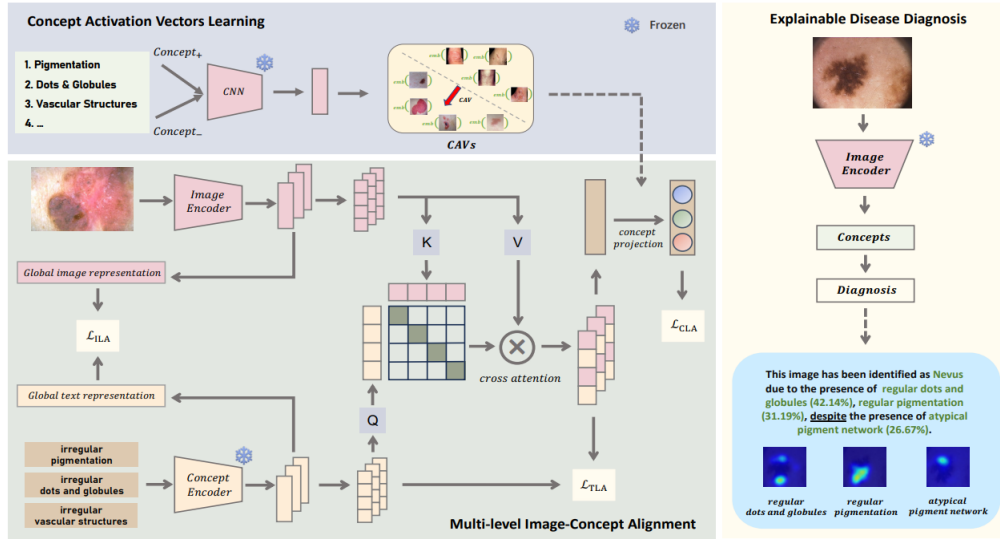


Figure 2: The overall pipeline of our proposed framework.

Figure 2.4: Overview of the MICA framework [7] for explainable skin lesion diagnosis. The model aligns visual and textual concepts at multiple semantic levels—image, token, and concept—to ensure global, localized, and categorical interpretability. Concept activation vectors are first extracted using a frozen CNN and then aligned with visual features using a cross-attention mechanism. During inference, the model outputs concept contributions and generates text- and region-based explanations for clinical interpretability. Source: [7].

Recent developments in Vision-Language Models (VLMs) have enabled the integration of rich semantic information into concept-based modeling. While VLM-based CBMs offer an interpretable approach to model development with strong predictive performance, they heavily depend on curated concept annotations and filtering processes to define the concept space. Furthermore, the reliability of concept predictions in such models may be questionable, as they critically assume that the visual latent representations of the backbone model accurately capture and align with the predefined textual concepts.

## 2.4 Summary and Research Objectives

Despite promising advancements in concept-based explainability methods for medical imaging, substantial challenges remain in effectively integrating clinical knowledge into these approaches. Key issues include:

- Difficulty in defining clinically meaningful and interpretable concepts.

- The extensive annotation effort required to train concept-supervised models.
- Reliance on ambiguous heatmaps for concept localization, limiting spatial interpretability.
- Concerns regarding the faithfulness of vision models to their corresponding textual concept representations.

Addressing these challenges remains an important obstacle to the widespread clinical adoption and integration of concept-based explainability methods. Given the strength of deep vision models in capturing subtle visual cues, we focus exclusively on the visual modality to preserve this potential and to build unsupervised concept-based explainability within the visual domain, bypassing the reliance on textual or manually defined concepts. Moreover, we aim to provide not only concept-based explanations but also spatial localization of concepts through patch-level representations. Our research specifically addresses these challenges by presenting an integrative framework that combines unsupervised concept discovery through non-negative matrix factorization with structured modelling of inter-concept relationships via graph-based representations. To this end, we propose building end-to-end concept graphs that serve as inherently interpretable models, enabling both concept reasoning and localization.

The primary goal of this research is to move beyond direct causal relationships between a single visual concept-output pair and attain a better understanding of the model’s decision-making process by incorporating relationships between concepts. This can be particularly important for applications such as medical imaging analysis, where understanding how the presence of different high-level semantic features collectively influences the prediction can be more significant than just identifying individual key features. To achieve this, we benefit from recent research to offer a unified approach to avoid common pitfalls in concept discovery. Our integrated framework pursues three key objectives: (a) uncover essential visual concepts via NMF in unsupervised manner, (b) establish data-driven associations between visual concept-output pairs through a refined graph representation. To fulfil these objectives, we adopt an additive concept representation and model inter-concept relationships within structural concept graphs. In particular, we take advantage of the non-negative additive linear property inherent in NMF to map the entire image into a structural concept graph, allowing for a comprehensive capture of the underlying data distribution. By employing an additive concept pooling strategy, our model makes it possible to analyse:

- How much each concept contributes to the final prediction.
- How much each patch influences the prediction, enabling localization of the most important regions.
- What concept-related information is present at each patch location.

In summary, this work introduces an automated, visually grounded concept-based explanation method that leverages robust and meaningful concept representations to enhance the transparency and interpretability of model predictions, aiming to inspire future works in the XAI field and foster greater trust in AI-assisted decision-making.

---

# Chapter 3

## Technical Background

### 3.1 Deep Learning

This chapter reviews the fundamental components of deep learning that are essential for the development of our proposed method. The objective is to familiarize the reader with the underlying deep learning paradigms that form the basis of the methodology. For comprehensive coverage of deep learning paradigms, readers may refer to the foundational textbook Deep Learning by Goodfellow, Bengio, and Courville [32], as well as the more recent Deep Learning: Foundations and Concepts by Christopher and Hugh Bishop [8], alongside the main reference papers cited throughout this section.

#### 3.1.1 Neural Networks

##### Fully Connected Layers

Fundamentally, neural networks function as a transformation operation that project input data to the given output space. The transformation of data is contingent upon the activity of neurons, which are responsible for the propagation of data to its new state. A neuron, which is the fundamental components of neural networks, is a function  $\eta : \mathbb{R}^d \rightarrow \mathbb{R}$  and is parametrised by a set of parameters such as weights and biases. These parameters are collectively referred to as  $\theta$  and are pivotal components of neural networks across various applications. In general, the function  $\eta$  can be described as follows:

$$\eta(x; \theta) := \sigma(w^\top x + b),$$

where  $w \in \mathbb{R}^d$  is the weight vector,  $b \in \mathbb{R}$  is the bias term,  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is non-linear activation function [108].

Neurons aggregate input signals in a linear manner, as defined by the  $\theta$  parameters. Subsequently, they apply a non-linear activation function  $\sigma$  (e.g., ReLU [77]) in an element-wise manner to produce the output. The employment of linear aggregation in



conjunction with non-linear transformation enables models to learn complex relationships between input and output signals.

In a manner analogous to the natural brain from which it was inspired, a neural network architecture is essentially made up of a large group of neurons that are interconnected by multiple layers. Stacking a large number of layers helps us create a large and powerful neural network architecture, which we refer to as deep neural networks. A fully connected feedforward neural network (FCNN) with  $L$  layers is characterized by the parameter set  $\theta$ , can be defined as:

$$f(x; \theta) := (\eta^{(L)} \circ \dots \circ \eta^{(1)})(x),$$

here  $\eta^{(i)}(a; \theta^{(i)}) = \sigma(W_{(i)}^\top a + b_{(i)})$  represents the transformation applied at the  $i$ -th layer,  $\theta^{(i)} = \{W_{(i)}, b_{(i)}\}$  holds the set of learnable parameters for that layer,  $a$  indicates the output from the preceding layer, and  $\sigma$  is a non-linear activation function.

The *final layer*, also known as the output layer, in a neural network is the last component in the architecture, allowing the network to produce predictions, while its structure is dependent on the nature of the prediction task. For continuous output predictions, the output layer constitutes a linear layer with no activation. For classification tasks, the output layer employs *sigmoid* and *softmax* activation functions to generate class probabilities, for binary classification and multi-class classification, respectively.

### 3.1.2 Model Training

Principles of model learning methods are firmly embedded in deep learning methodologies. These principles not only allow us to teach our model, but also ensure that the input-output mapping is robust and sensible. In this section, we will discuss some of the strategies essential for model training.

#### Supervised Learning

In this subsection, we will briefly discuss how to train our neural network to learn the desired data representation using supervised learning. Within the context of supervised learning, models are trained using labelled input-output pairs, enabling them to infer patterns and generate predictions or categorize new, unseen data. Denoting the measurable input and output spaces by  $\mathcal{X} \subset \mathbb{R}^d$  and  $\mathcal{Y} \subset \mathbb{R}$ , respectively, the dataset  $\mathcal{D}$  can be represented in a supervised learning setting as a group of annotated input-output pairs:

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n.$$

Within the framework statistical learning, the objective is to approximate the underlying stochastic relationship between an input and label pairs. This relationship is expressed as the expected value of  $y$  conditioned on  $x$ , with respect to the joint probability distribution  $P_{x,y}$  defined over input-output spaces. To approximate this relationship, we define a hypothesis space  $\mathcal{F}$  with a predictor model  $f : \mathcal{X} \rightarrow \mathcal{Y}$  and a loss function  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  serves as a quantitative analysis of similarity between the output and the actual target value.

In a binary image classification task for example,  $X$  represents the samples of input images and  $\mathcal{Y} = \{+1, -1\}$  represent positive and negative labels. The binary cross-entropy loss can be employed as a potential loss function for this task. As the original distribution

of  $P_{x,y}$  is unknown, the learning objective is approximated using the empirical distribution derived from the dataset. Given a training dataset  $\mathcal{D}$ , a deep learning predictor  $f$ , and a loss function  $\ell$ , the empirical risk can be expressed as:

$$\mathcal{L}(f) := \frac{1}{n} \sum_{i=1}^n \ell(f(x_i; \theta), y_i)$$

In supervised learning, the ultimate training aim is to ascertain the ideal parameters  $\theta^*$  that minimize the expected loss across the training samples, formulated as:

$$\theta^* = \arg \min_{\theta} \mathcal{L}(f)$$

### Loss Function

The loss function is an important component of model learning, and selecting an appropriate loss function is a critical decision when applying deep learning to a specific task. The most commonly employed loss functions in supervised learning are the mean squared error (MSE) and cross-entropy loss, each guiding the optimization process by computing the disparity between predicted and true labels. MSE is used for regression tasks, while cross-entropy loss is a common choice for classification tasks. As the scope of this work involves multiclass classification problems, cross-entropy loss is an appropriate choice for this setting. For the true label  $y \in \{1, \dots, C\}$  and the predicted class probabilities  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_C$ , the cross-entropy loss for multiclass setting can be defined by:

$$\ell = \mathcal{H}(y, \hat{y}) = - \sum_{c=1}^C y_c \log(\hat{y}_c),$$

where  $y_c \in \{0, 1\}$  (one-hot encoded) is the original label, and  $\hat{y}_c$  is the predicted probability for class  $c$ .

### Regularization

Minimizing  $\mathcal{L}(f)$  has two main pitfalls. The first is that the problem is usually non-convex, except in simple cases like linear regression. Second, empirical risk minimization does not guarantee low error on unseen data, which often leads to overfitting. This phenomenon limits the ability of models to generalize and requires regularization.

To remedy this, the regularized empirical loss function incorporates a penalty term into the original empirical loss. This penalty term directly affects the model parameters by constraining their values during optimization, thereby preventing overfitting and making the model more robust to unseen data. A general regularized empirical loss term can be formulated with the following equation:

$$\mathcal{L}_{\text{reg}}(f) := \mathcal{L}(f) + \lambda \Omega(\theta),$$

where  $\Omega(\theta)$  represents the regularization function (e.g.,  $\|\theta\|_2^2$  for L2,  $\|\theta\|_1$  for L1) and  $\lambda \geq 0$  is the regularization hyperparameter controlling the penalty strength.

### Dropout

Dropout is another powerful regularization method widely used in neural networks for robust training. It randomly sets a portion of the neurons' activations in the selected

layer to zero (or drops them out) with a certain probability  $p$ . Dropout is only applied during the learning phase to encourage the model to reduce its reliance on any single neuron or specific patterns in the feature space, thereby improving generalization.

### Batch Normalization

Another important component among training methodologies is the batch normalization [47] technique, which is designed to stabilize and accelerate the training of neural networks by tackling the internal covariance shift. The internal covariance shift arises from how the parameters are initialized and how the input distributions to each layer shift during training, which can hinder learning efficiency and lead to slower or unstable training. To address this, Batch Normalization standardises the input features of every mini-batch to zero mean and unit variance, subsequently scaling and shifting them using the learnable parameters  $\gamma$  and  $\beta$ . Specifically, for an input feature  $x$  in a mini-batch of  $B$ , this method is defined as:

$$\text{BN}_{\gamma, \beta}(x) = \alpha \left( \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \beta,$$

where  $\gamma$  and  $\beta$  are trainable parameters,  $\mu_B$  and  $\sigma_B^2$  represent the mean and variance of a batch data, and  $\epsilon$  denotes a small constant value for numerical stability.

In short, Batch Normalization speeds up training, mitigates the vanishing/exploding gradients problem, and is widely used in modern neural networks.

### Residual connections

Vanishing gradients are a common issue in neural networks that can hinder or destabilize the learning process [5]. Due to the repeated multiplication of gradients during backpropagation, the values can either shrink to near zero (vanishing) or grow excessively large (exploding), particularly as network depth increases. Residual connections, also known as skip connections, were introduced to address this issue and facilitate the training of deeper networks [40]. These connections have been instrumental in the development of state-of-the-art deep learning architectures, especially those of vision models [40].

Residual connection mechanism work by creating a shortcut that skips at least one layer within a given architectural block, directly adding the input to the output of that block. This facilitates better gradient flow throughout the network, helping to prevent performance degradation in very deep models. A simple residual connection works by feeding the activations from a previous layer directly into the next layer, which can be defined by:

$$r(x; \theta) = x + \eta(x; \theta),$$

here  $\eta(x; \theta)$  shows the transformation function (e.g., a sequence of layers) with parameters  $\theta$ ,  $x$  is the input, and  $r(x; \theta)$  is the output of the residual block.

### Optimization Method

In empirical risk minimization, models are typically parametrized by a set of learnable parameters  $\theta \in \Theta$  and a learning process that optimize such parameters to minimize the empirical risk. Among such methods used for optimization in deep learning, Stochastic Gradient Descent (SGD) [60] is a fundamental choice. This method updates  $\theta$  using

gradients computed on small, randomly selected mini-batches of data as opposed to computing gradients over the entire dataset, which is computationally expensive. This makes SGD significantly more efficient for large-scale models. Given that  $\ell$  defines the loss function,  $\alpha$  the learning rate, and  $B = \{(x_i, y_i)\}_{i=1}^{|B|} \subset \mathcal{D}$  a mini-batch of samples, the parameters are updated as:

$$\theta_{t+1} = \theta_t - \alpha \sum_{i=1}^{|B|} \nabla_{\theta} \ell(f(x_i; \theta_t), y_i),$$

here,  $\nabla_{\theta} \ell(\cdot)$  denotes the gradient of the loss function  $\ell$  with respect to the learnable parameters  $\theta$ , and  $\alpha$  functions as the learning rate, adjusting the size of each step taken in the opposite direction of the gradient during optimization.

This iterative approach not only enables scalable training on large datasets but also introduces beneficial noise that can help escape local minima in non-convex loss landscapes, making it particularly effective for deep neural networks. Similar in core principles, other optimization algorithms such as Adam [54] and RMSprop [105] are also commonly used during training.

### 3.1.3 Convolutional Neural Networks

Introduction of deep Convolutional Neural Networks for ILSVRC in 2012, marked a major inflection point for deep learning in the field of computer vision [58]. Inspired by the human visual system, CNN is a class of DL that uses convolutional operations to automatically extract and recognize patterns from grid-structured data, particularly in images.

Similar to deep neural networks, CNN shares similar structures, consisting of multiple layers of fully connected layers. The main building blocks of CNN are made up convolution and pooling layers. Modern CNN models include multiple blocks of convolutional layers, typically followed by a pooling operation and a non-linear activation function.

#### Convolution layer

Convolution layers are the layers in which the convolution operation is applied to the image. More specifically, these layers apply learnable filters over local regions of the image, capturing spatial features such as edges or textures. By stacking series of convolution layers, CNN can learn more abstract and global features such as object.

For a single-channel case, the convolution between the input image  $x$  and filter  $w$ , defined by  $\theta = \{w, b\}$ , is formulated as:

$$(x * w)_{i,j} = \sum_{u=0}^{k_h-1} \sum_{v=0}^{k_w-1} x_{i+u, j+v} \cdot w_{u,v} + b,$$

where  $*$  denotes convolution operation,  $w$  is the convolutional filter (kernel) of size  $k_h \times k_w$ , and  $(u, v)$  iterate over the height and width of the kernel. For multi-channel inputs, each channel is convolved independently, and the results are summed.

Once the convolution operation is applied, the output is subsequently transformed with a non-linear activation function, such as ReLU, to add non-linearity into the features and enable richer representations. This process produces what is referred to as a *feature*

or *activation map*. The dimensions of the resulting feature map is  $W \times H \times C$ , where  $C$  denotes the number of channels (i.e., learnable filters), and  $W$  and  $H$  represent the spatial width and height, respectively.

### Pooling layer

Decreasing the dimensions of feature maps in CNNs is key to making the model efficient, generalizable, and better at recognizing complex patterns. This is done through pooling layers, which shrink the width and height of feature maps, reducing data size, speeding up training, and lowering memory use. It also simplifies the model by decreasing parameters, which helps prevent overfitting. Additionally, this reduction enlarges the receptive field of deeper layers, allowing neurons in deeper layers to “see” bigger portions of the input and capture more complex features. Pooling also provides translational invariance, making the model less sensitive to small shifts in the input. Overall, reducing spatial dimensions helps CNNs remain efficient while learning richer, hierarchical representations.

The most widely applied pooling methods are *Max Pooling* and *Average Pooling*. To reduce the dimensionality, *Max Pooling* picks the highest value within a given region, whereas the *Average Pooling* method calculates the mean value over this region. Another commonly used pooling strategy is *Adaptive Average Pooling* [63]. The *Adaptive Average Pooling* method dynamically adjusts the pooling windows to produce an output of a fixed size relative to the input dimensions, making it particularly useful for handling inputs of varying sizes.

### 3.1.4 Graph Neural Networks

Graph Neural Networks (GNNs) [55] are a special type of neural network developed to operate on non-Euclidean data structures, such as graphs. Contrary to conventional neural network models, which expect inputs to be arranged on a regular grid structure (e.g., image pixels or word tokens), GNNs can learn representations from arbitrarily structured data by directly leveraging the graph topology and the features of nodes and edges.

GNNs have become the most common choice of models in deep learning on graphs, thanks to their effectiveness in capturing local and global graph dependencies via *message passing* and their permutation-invariant nature, making them suitable for relational data. These abilities makes them a powerful tool for formulating real-world problems such as social networks, citation networks, molecular network, and knowledge graphs [27].

Formally, a graph can be represented as  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where the set of  $N = |\mathcal{V}|$  nodes is defined by  $\mathcal{V}$ , and the set of edges is defined by  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ , possibly including self-loops. Each node  $v \in \mathcal{V}$  can have a feature vector  $x_v \in \mathbb{R}^d$ , and each edge  $(u, v) \in \mathcal{E}$  can carry an edge feature  $e_{uv}$ . At the core of GNNs, lies the *message passing* framework, in which information is iteratively aggregated from each node’s local neighbourhood. A single GNN layer can typically be defined by the following formulation:

$$h_v^{(k)} = \gamma^{(k)} \left( h_v^{(k-1)}, \oplus_{u \in \mathcal{N}(v)} \phi^{(k)}(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}) \right),$$

where  $h_v^{(k)}$  indicates the intermediate embedding of node  $v$  at layer  $k$ ,  $\mathcal{N}(v)$  is the set of neighbours of node  $v$ ,  $\phi^{(k)}$  is the message function that computes messages from neighbouring nodes (possibly using edge features too),  $\oplus$  denotes a permutation-invariant

aggregation function (e.g., mean or maximum), and  $\gamma^{(k)}$  is the update function, often implemented as a multilayer perceptron (FCNN) or a simple non-linear transformation. This iterative process updates the each node's embedding by combining its own current state with information from its neighbours. The initial node features are denoted as  $h_v^{(0)} = x_v$ .

### Graph Attention Networks

Graph Attention Networks [110] are a powerful extension of GNNs that incorporate the self-attention mechanism into the message passing process. Unlike traditional GNNs, which aggregate messages from neighbouring nodes using fixed or uniform weights (e.g., based on degree), GATs learn to assign dynamic, data-dependent importance scores to each neighbour, also known as attention scores. This enables the graph to focus more on relevant neighbours and less on noisy or irrelevant nodes. These coefficients are computed as follows:

$$\alpha_{vu}^{(k)} = \frac{\exp \left( \text{LeakyReLU} \left( a^\top [\mathbf{W}h_v^{(k-1)} \parallel \mathbf{W}h_u^{(k-1)}] \right) \right)}{\sum_{w \in \mathcal{N}(v)} \exp \left( \text{LeakyReLU} \left( a^\top [\mathbf{W}h_v^{(k-1)} \parallel \mathbf{W}h_w^{(k-1)}] \right) \right)},$$

where  $h_v^{(k-1)}$  and  $h_u^{(k-1)}$  denote the feature vectors of nodes  $v$  and  $u$ , respectively, at layer  $(k-1)$ , the notation  $\mathcal{N}(v)$  indicates the neighbouring nodes of  $v$ ,  $\mathbf{W}$  and  $a$  are a learnable weight matrix and attention vector, respectively,  $\parallel$  represents the concatenation operation, and LeakyReLU is a non-linear activation function.

To compute  $\alpha_{vu}^{(k)}$ , each node's input feature vector  $h_v^{(k-1)}$  is first linearly transformed using a shared weight matrix  $\mathbf{W}$ . For each node  $v$  and its neighbour  $u$ , an unnormalized attention score is then computed by applying a non-linear activation function (LeakyReLU [68]) to the concatenated transformed features. These scores are subsequently normalized across all neighbours of  $v$  using the softmax function to produce the final attention coefficients  $\alpha_{vu}^{(k)}$ . Once attention scores are computed, node representations are updated via a weighted sum of their neighbours:

$$h_v^{(k)} = \sigma \left( \sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(k)} \cdot \mathbf{W}h_u^{(k-1)} \right),$$

where  $\sigma$  is a non-linear activation function, such as ELU [16] or ReLU.

In some cases, optional *edge weights* can be incorporated into the attention mechanism [103]. These are not used to compute the attention scores directly but are instead applied after the softmax normalisation as multiplicative factors, allowing the integration of prior knowledge or structural importance encoded in edge features. The updated final aggregation thus becomes:

$$h_v^{(k)} = \sigma \left( \sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(k)} \cdot \text{edge\_weight}_{vu} \cdot \mathbf{W}h_u^{(k-1)} \right),$$

where  $\text{edge\_weight}_{vu}$  is a scalar weight modulating the influence of neighbour  $u$ 's message to node  $v$ . This approach enables GATs to be extended to cases where the graph structure contains meaningful edge-level information (e.g., similarity scores).

### Multi-head attention

Multi-head attention, originally introduced by Vaswani et al. [109] as a key component of the Transformer architecture for simultaneous sequence modelling in machine translation, has since been widely adopted in other architectures, such as GAT. This mechanism facilitates the extraction of diverse contextual features from different parts of the input simultaneously using distinct learnable parameter sets.

Rather than learning a single attention score, the input data is distributed over  $n$  attention heads, each with its own set of learnable query, key, and value projection parameters. For each:

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)$$

where  $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_k}$  are learnable query, key, and value parameters, and  $d_k = d/n$ . Scaled dot-product attention is used within each head  $i \in \{1, \dots, n\}$ :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

At the final step, the individual outputs from all attention heads are combined via concatenation and transformed by a subsequent linear layer, thereby concluding the multi-head attention process:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_n)\mathbf{W}^O$$

with  $\mathbf{W}^O \in \mathbb{R}^{d \times d}$  as a learnable output projection. This mechanism allows each head to capture distinct aspects of the input—such as syntactic structure, positional cues, or long-range dependencies—resulting in more robust and expressive representations.

Multi-head attention has since been integrated into various architectures beyond NLP, including computer vision (e.g., Vision Transformers) and graph neural networks. In particular, it has been adapted in GATs, where each head learns to focus on different local substructures in the neighbourhood of a node [110]. In GATs, the input sequence is replaced by a graph structure, and attention is computed over the local neighbourhood of each node. For each attention head  $i \in \{1, \dots, n\}$ , the node representation is updated as:

$$h_v^{(k,i)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(k,i)} \cdot \mathbf{W}^{(k,i)} h_u^{(k-1)}\right),$$

where  $\alpha_{vu}^{(k,i)}$  is the attention coefficient between nodes  $v$  and  $u$  for head  $i$ , and  $\mathbf{W}^{(k,i)}$  is a learnable projection at  $k$ -th layer for head  $i$ . The final node representation at any given layer is then obtained by either:

$$h_v^k = \parallel_{i=1}^n h_v^{(k,i)} \quad \text{or} \quad h_v^k = \frac{1}{n} \sum_{i=1}^n h_v^{(k,i)}$$

depending on whether concatenation or averaging is used (typically, used in the final layer). This allows the model to learn different substructures or semantic roles in the neighbourhood of a node, enhancing expressiveness and robustness. Consequently, GATs can learn richer and more context-aware node representations, while reducing overfitting and stabilising training via diverse attention heads.

## Graph Normalization

Unlike Batch Normalization, which operates across samples in a batch, Graph Normalization [10] performs normalisation within each individual graph. This intra-graph approach respects the variable sizes and topologies of graphs, contributing to more stable training and improved generalisation, particularly for graph-level prediction tasks.

The normalisation is performed as follows:

$$\mathbf{h}' = \frac{\mathbf{h} - \alpha \odot \mathbb{E}[\mathbf{h}]}{\sqrt{\text{Var}[\mathbf{h} - \alpha \odot \mathbb{E}[\mathbf{h}]] + \epsilon}} \odot \gamma + \beta,$$

where  $\mathbf{h}$  is the node feature,  $\mathbb{E}[\mathbf{h}]$  denotes the mean of all node features in the graph,  $\alpha$  is a learnable scalar that adjusts the extent of centring, and  $\gamma, \beta$  are learnable affine transformation parameters. The operator  $\odot$  denotes element-wise multiplication, and  $\epsilon$  ensures numerical stability.

The studies have shown that integrating Graph Normalization into GNN architectures leads to consistent improvements in both performance and convergence speed across a range of graph classification tasks.

## 3.2 Matrix Decomposition

Matrix decomposition techniques are widely used in machine learning to uncover latent structures within high-dimensional representations. Methods such as Principal Component Analysis and Non-negative Matrix Factorisation help identify dominant patterns—often referred to as principal components or latent concepts—that can be leveraged for interpretation, compression, or clustering. In the context of neural networks, these decompositions are particularly valuable for mapping complex activation spaces into more interpretable forms, enabling concept-based analysis and explanation.

### 3.2.1 Non-negative Matrix Factorization

Non-negative Matrix Factorisation is a dimensionality reduction technique that decomposes a non-negative data matrix into the product of two lower-rank non-negative matrices [61]. For a given non-negative activation matrix  $A \in \mathbb{R}_{\geq 0}^{n \times p}$ , NMF seeks to factorize  $A$  into two lower-rank non-negative matrices:

$$A \approx \mathbf{U}\mathbf{V}^\top,$$

where  $\mathbf{U} \in \mathbb{R}_{\geq 0}^{n \times r}$  and  $\mathbf{V} \in \mathbb{R}_{\geq 0}^{p \times r}$  for some lower rank  $r \ll \min(n, p)$ , and  $n$  denotes the number of samples (e.g., images) and  $p$  the size of feature vector (e.g., neuron activations in a given layer of a neural network).

In the context of interpretability in deep learning, NMF has been effectively used to identify a compact and interpretable set of latent directions in the activation space, which acts as a concept basis CAV in concept-based explanations [121]. Here, the matrix  $\mathbf{V}$  contains  $r$  such CAVs, each representing a coherent direction in the feature space that captures a recurring pattern or concept across the dataset. Meanwhile,  $\mathbf{U}$  expresses the activations of each input sample in terms of its contributions from the discovered concepts.



This decomposition yields a semantic basis in which each image can be understood as a non-negative linear combination of concepts. The non-negativity constraint not only ensures interpretability—since negative contributions are not allowed—but also encourages sparsity in both  $U$  and  $V$ . This often leads to disentangled and localised representations where only a few concepts are active per image, and each concept is defined by a selective subset of features.

### 3.3 Evaluation Metrics

To assess the effectiveness of our approach, we employ both standard classification performance metrics and specialized metrics tailored for evaluating the quality and faithfulness of extracted visual concepts.

#### 3.3.1 Model Performance

We evaluate the predictive performance of our trained models using common classification metrics, such as Accuracy, F1 Score, and the Area Under the Receiver Operating Characteristic Curve (AUC), in line with previous works [7, 85, 44].

#### 3.3.2 Concept Quality Evaluation

To evaluate the quality of the extracted concepts, we adopt a set of five quantitative metrics used in prior works [21, 121]. These metrics capture the desired properties of a concept representation in terms of *reconstruction error*, *sparsity*, *stability*, and *distribution alignment*. The metrics include:

- **Relative  $\ell_2$  Distance ( $\Downarrow$ ):** Measures reconstruction quality by comparing the activation matrix  $A$  with its reconstruction  $UV^\top$ . Lower values indicate more faithful reconstruction.
- **Sparsity ( $\Uparrow$ ):** Quantifies how sparse the concept encoding  $U$  is, defined as  $\|u\|_0/k$ , where  $\|u\|_0$  is the number of non-zero concept activations in the encoding vector  $u \in \mathbb{R}^k$ , and  $k$  is the total number of concepts. This metric reflects the fraction of concepts active for each patch—lower values indicate that fewer concepts are used, promoting more interpretable and disentangled representations.
- **Stability ( $\Downarrow$ ):** Assesses how reliably the concept extraction method recovers similar concepts across different subsets of data. To evaluate this, concept extraction method is applied on multiple K-fold splits and the resulting concepts are matched using the Hungarian loss function. The stability score is then computed by measuring the average cosine similarity between corresponding Concept Activation Vectors (CAVs) across folds. A lower score indicates greater consistency, implying that the method captures robust, reproducible concepts rather than artifacts specific to a particular subset.
- **(FID) ( $\Downarrow$ ):** FID quantifies the discrepancy between the distributions of original and reconstructed activations using the *1-Wasserstein distance* [111]. A lower FID indicates that the learned concept representations maintain the overall structure and diversity of the original data distribution.

- **Out-of-Distribution (OOD) Score ( $\downarrow$ ):** The OOD score assesses how well reconstructed samples remain within the data distribution. It relies on *Deep-KNN* [101] score, which measures the proximity of each reconstructed point to its nearest neighbours in the original dataset. Lower scores suggest reconstructions that are more consistent with the support of the original data, indicating better in-distribution alignment.

These metrics enable a multifaceted evaluation of concept representations beyond mere visual coherence, ensuring they are both interpretable and structurally sound within the model’s learned space.

### 3.3.3 Faithfulness Evaluation

To validate whether the extracted concepts genuinely influence the model’s decision-making, we employ **insertion** and **deletion** tests, as described in prior works [31, 121, 51]. These tests involve perturbing the input by either inserting or removing concept-relevant regions and observing the effect on the model’s output:

- **Insertion Test:** Measures the change in model confidence when concept-relevant regions are gradually added back to a blurred image. A faithful concept will significantly increase confidence upon insertion.
- **Deletion Test:** Measures the drop in confidence when concept-relevant regions are progressively removed. A large drop indicates that the concept was crucial for the original prediction.

# Chapter 4

## Methodology

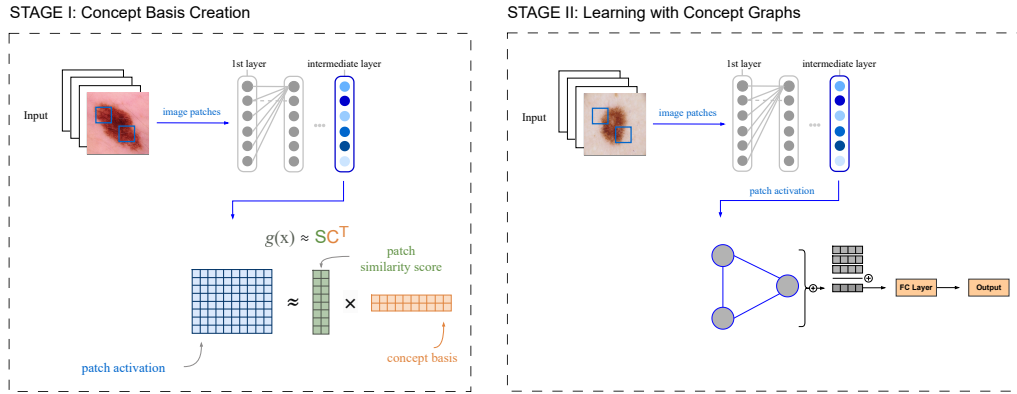


Figure 4.1: Overview of our proposed visually grounded concept-based framework. **i)** Input images are divided into random crops to generate image patches, which are passed through a frozen intermediate layer of a vision backbone. From the resulting patch activations, a concept basis is extracted using non-negative matrix factorisation (NMF), providing interpretable visual concepts independent of class labels. **ii)** In the second stage, patch activations are systematically aggregated into concept nodes based on the concept basis via a patch pooling mechanism, forming a concept graph. This graph is processed by a Graph Attention Network to model inter-concept relationships. The framework is trained in an ante-hoc manner, enabling interpretable, visually grounded predictions.

In this thesis, we propose a two-phase framework for learning interpretable, concept-based visual representations using ante-hoc training with graph neural networks (GNNs). Our approach first extracts a high-level visual concept basis from the dataset, then formulates concept graphs and performs end-to-end training that incorporates these human-understandable concepts. The methodology consists of two main stages: *i) Concept Basis Creation*, and *ii) Learning with Concept Graphs* (Figure 4.1). In the following sections, we detail how each of these stages is formulated, discuss the associated design

choices, and explain how the resulting concept graphs can be interpreted to provide insight into the model’s decision-making process.

## 4.1 Concept Basis Generation

The automatic extraction of human-understandable visual concepts from the dataset forms a foundational step in our proposed approach, particularly within frameworks designed to enhance the interpretability of deep learning models. This section presents the concept basis generation step in our implementation, which utilizes non-negative matrix factorization (NMF), borrowed from prior work [25, 121]. This stage of the pipeline enables us to represent and interpret images using an interpretable set of concepts denoted by  $\mathcal{K} = \{k_1, k_2, \dots, k_r\}$ , where  $r$  corresponds to the number of learned concepts in the basis.

We begin concept generation by collecting the images for which we aim to compute a concept basis representing existing high-level semantic features. Rather than focusing on a specific class, we consider the entire set of images  $\mathcal{X} = \{x_i\}_{i=1}^n$  from the dataset, independent of their class labels. This enables us to learn a class-agnostic concept basis that captures the model’s internal representations across all classes collectively.

Concept-based explanation methods fundamentally rely on detecting and explaining local high-level semantic features that exist within the dataset. To achieve this, we first create a pool of candidate concepts by extracting image sub-regions. While some techniques use segmentation masks and inpainting to generate candidate concepts, these approaches can introduce artefacts in latent space. Instead, we leverage the fact that modern deep networks are commonly trained using data augmentations such as RandAugment, Mixup, and CutMix, which incorporate cropping operations. Following this insight, we employ a straightforward cropping and resizing procedure, denoted by  $\pi(\cdot)$ , to extract image patches, as proposed in [25].

From each image  $x_i \in \mathcal{X}$ , we extract multiple sub-regions by applying  $\pi(\cdot)$ , resulting in an auxiliary dataset  $\mathbf{X}' \in \mathbb{R}^{l \times d}$ , where each element  $\mathbf{X}'_i = \pi(x_i)$  is a crop. Using cropping instead of segmentation and inpainting avoids artefacts while aligning with common data augmentation strategies. Next, the cropped images  $\mathbf{X}'$  are passed through the intermediate layer of deep neural network  $g(\cdot) \subseteq \mathbb{R}^p$  to obtain latent activations:

$$\mathbf{A} = g(\mathbf{X}') \in \mathbb{R}^{n \times p}. \quad (4.1)$$

Note that, for convolutional neural networks, global average pooling is applied to convert spatial activations into vector form. To extract a low-dimensional concept basis, we apply NMF to decompose the activation matrix  $\mathbf{A} \in \mathbb{R}^{n \times p}$  into two non-negative matrices  $\mathbf{S} \in \mathbb{R}^{l \times r}$  and  $\mathbf{C} \in \mathbb{R}^{p \times r}$ , where  $r \ll \min(l, p)$ , by solving:

$$(\mathbf{S}, \mathbf{C}) = \arg \min_{\mathbf{S} \geq 0, \mathbf{C} \geq 0} \frac{1}{2} \|\mathbf{A} - \mathbf{S}\mathbf{C}^\top\|_F^2, \quad (4.2)$$

here  $\|\cdot\|_F$  denotes the Frobenius norm,  $\mathbf{C}$  contains the concept-basis matrix (CAV), while  $\mathbf{S}$  provides the coefficients that represent each input crop as a combination of concepts. The non-negativity constraints encourage sparsity, yielding disentangled concepts (in  $\mathbf{C}$ ) and compact representations (in  $\mathbf{S}$ ) aligned with ReLU activations, which aids interpretability and missing data imputation [88].

Utilizing matrix factorization, the activation of each input crop  $\mathbf{X}'_i$  can be reconstructed as a weighted combination of concepts, given by:

$$A_i = \sum_{j=1}^r \mathbf{S}_{i,j} \mathbf{C}_j^\top. \quad (4.3)$$

This additive, parts-based decomposition enhances the interpretability of the model’s internal representations, and give us ability to represent new crops using computed concept basis  $\mathbf{C}$ . In the Section 4.2, we will detail how we employ the parts-based decomposition nature to formulate the concept graphs, and train our model.

## 4.2 Learning with Concept Graphs

Formulating concept graphs lies at the core of our proposed approach. This design draws inspiration from Concept Bottleneck Models, where the final layer consists of a set of interpretable concept activations that directly influence model predictions. However, in contrast to CBMs, we ground our node features entirely within the visual modality, bypassing the reliance on textual or manually defined concepts. To this end, we formulate concept graphs using a Graph Attention Network in an *ante-hoc* fashion. GATs are particularly well-suited for this task, as they allow for rich representation learning while explicitly modelling interactions between concepts through learned attention mechanisms. Representing an image  $I$  as a graph, where each node corresponds to a specific visual concept, constitutes a fundamental component of our methodology. In the following, we outline how this graph construction is achieved and present the general formulation of our GAT-based framework.

Given  $r$  visual concepts from the set of concepts  $\mathcal{K}$ , we define a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{H}, \mathbf{E}_w)$  consisting of  $r = |\mathcal{V}|$  nodes. Each node  $v \in \mathcal{V}$  has a feature vector  $\mathbf{h}_v \in \mathbb{R}^p$ , and the collective node features are represented by the matrix  $\mathbf{H} \in \mathbb{R}^{r \times p}$ . The edge set  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  specifies the connectivity of the graph and each edge is represented by  $(u, v) \in \mathcal{E}$ . While we do not introduce explicit edge features, we encode prior knowledge about inter-concept relationships using an edge weight matrix  $\mathbf{E}_w \in [0, 1]^{r \times r}$ . Each entry  $(\mathbf{E}_w)_{uv}$  reflects the strength or confidence of the connection from node  $u$  to node  $v$ , with a value of zero indicating no connection. These edge weights are incorporated into the attention mechanism during message passing within the GAT, enabling the model to softly integrate concept-specific priors in a learnable and flexible manner.

### 4.2.1 Concept Graph Representation

One of the key limitations of existing ante-hoc explainability methods is their reliance on heatmaps for concept localisation, which often produce ambiguous and hard-to-interpret results. Moreover, as discussed in Section 2, many existing approaches require manual intervention during the concept building stage [28], and often lead to sub-optimal concept representations [31]. Our goal is to develop an automated, visually grounded concept-based explanation method with robust and meaningful concept representations.

As discussed in Equation 4.3, the part-based nature of Non-negative Matrix Factorisation (NMF) allows each input crop to be represented as a weighted combination of visual concepts. We leverage this property to construct concept graphs that can capture the structural properties of the underlying visual data distribution.

The central question is: how can an entire image  $x_i$  be meaningfully represented as a concept graph? The most straightforward approach would be to directly map image-

level activations to graph nodes. However, rather than relying on global image features, we aim to decompose the image into patches, enabling localised concept assignment within the image. This is motivated by the core principle of concept-based explanations: detecting and interpreting local high-level semantic features that emerge across different image regions.

To achieve this, we divide each image into patches using a systematic patch extraction process, similar to the cropping operation described in Section 4.1. However, unlike the random cropping used during concept basis generation, we apply a structured pooling with stride operation, denoted by  $\phi(\cdot)$ , inspired by convolutional pooling mechanisms. Specifically, given an image  $x_i \in \mathcal{X}$ , we extract multiple overlapping patches by applying  $\phi(\cdot)$ , resulting in a set of patches  $\mathbf{X} \in \mathbb{R}^{n \times d}$ , where  $n$  indicates the number of extracted patches. Throughout our experiments, we fix the pooling stride ratio at 0.5.

In practice, each patch may contain multiple visual concepts. Thanks to the additive structure of NMF, we can encode the degree to which each patch expresses different concepts and propagate this information to the corresponding concept nodes in the graph. Rather than applying simple discrete cropping, our pooling operation systematically slides over the image, enabling complete structural mapping of the image into its concept graph representation.

For each patch, we first estimate its concept composition by projecting its activation into the learned concept basis  $\mathbf{C}$  within the intermediate feature space. To that end, the following Non-Negative Least Squares (NNLS) problem is solved:

$$\mathbf{s}_i = \arg \min_{\mathbf{s} \geq 0} \frac{1}{2} \|\mathbf{g}(\mathbf{X}_i) - \mathbf{s}\mathbf{C}^\top\|_F^2 \quad (4.4)$$

where  $\mathbf{C}$  is the learned concept basis,  $\|\cdot\|_F^2$  indicates the Frobenius norm, and  $\mathbf{s}_i \in \mathbb{R}^r$  represents the concept composition coefficients for the patch  $i$ , with  $r = |\mathcal{K}|$ . The function  $\mathbf{g}(\cdot)$  denotes the same backbone vision model used consistently throughout the entire framework for extracting intermediate patch-level activations.

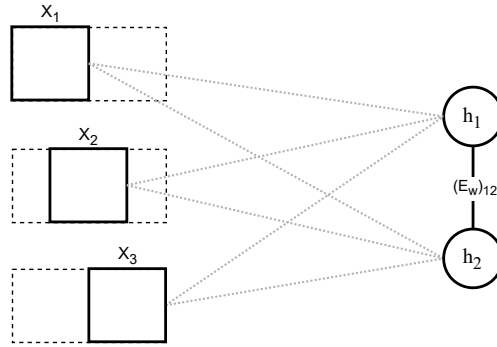


Figure 4.2: Illustration of the patch pooling mechanism. Patch-level activations are aggregated into concept nodes based on their concept composition coefficients  $\mathbf{s}$ . The process moves from left to right with stride, where the bold red rectangle indicates the current region of focus.

Next, each patch activation is weighted by its similarity score  $s_i$  and aggregated to the corresponding concept node. This process is repeated for all patches within the image, enabling the model to systematically aggregate local information into concept nodes. An

illustration of this patch pooling mechanism is shown in Figure 4.4. To prevent visual concepts that appear frequently within an image (e.g., background regions) from disproportionately influencing the final node representations, we apply a normalisation term during aggregation. Finally, we apply a soft non-linear activation function  $\sigma(\cdot)$ , chosen as the GELU activation [41], to promote rich, informative embeddings while maintaining alignment with the underlying activation distribution and supporting gradient flow.

Formally, the initial feature for each concept node is computed as:

$$\mathbf{h}_v^{(0)} = \sigma \left( \frac{\sum_{i=1}^n g(\mathbf{X}_i) \times s_{iv}}{1 + \sum_{i=1}^n s_{iv}} \right), \quad \text{for each concept } v \in \mathcal{K} \quad (4.5)$$

Here,  $s_{iv}$  refers to the scalar coefficient capturing the similarity of concept  $v$  to the embedding of patch  $i$ .

This normalised patch pooling scheme with soft non-linearity effectively integrates information from the entire image while allowing patches containing multiple concepts to contribute proportionally to their respective concept nodes.

Once the image information is encoded into the corresponding nodes of the concept graph, we further leverage the learned concept basis  $\mathbf{C}$  to guide the construction of edges between nodes. Given the concept basis matrix  $\mathbf{C} \in \mathbb{R}^{p \times r}$ , where each column  $\mathbf{c}_v \in \mathbb{R}^p$  represents the embedding of the  $v$ -th concept vector, we compute an edge weight matrix  $\mathbf{E}_w \in [0, 1]^{r \times r}$ . Each entry  $(\mathbf{E}_w)_{uv}$  encodes the cosine similarity between concept vectors  $\mathbf{c}_u$  and  $\mathbf{c}_v$ .

$$\cos(\mathbf{c}_u, \mathbf{c}_v) = \frac{\mathbf{c}_u \cdot \mathbf{c}_v}{\|\mathbf{c}_u\|_2 \|\mathbf{c}_v\|_2} \quad (4.6)$$

$$(\mathbf{E}_w)_{uv} = \cos(\mathbf{c}_u, \mathbf{c}_v) \quad (4.7)$$

This construction is designed to ensure that only semantically similar concepts exchange information during message passing, preventing information smoothing and enhancing the robustness of the learned concept graph.

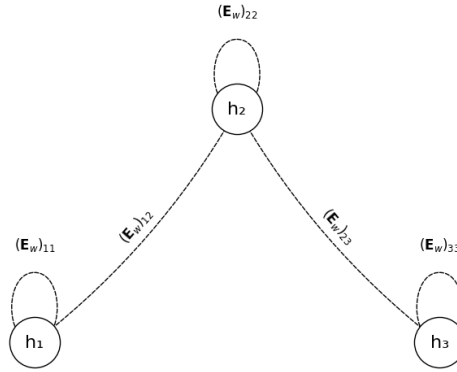


Figure 4.3: Concept graph representation example based on the edge weight matrix  $\mathbf{E}_w$  and feature vectors  $\mathbf{h}_v$ . Nodes represent concept embeddings, and edges encode pairwise cosine similarity. Edges with weak similarity are removed.

To illustrate an example structure of the concept graph, we provide a simplified example in Figure 4.3. Each node represents a distinct visual concept from the learned concept

basis, while edges are weighted according to the pairwise cosine similarity of the corresponding concept embeddings, as defined in Equation 4.7. To avoid message passing between weekly related concepts, edges with lower similarity are explicitly removed by setting the corresponding  $\mathbf{E}_w$  entries to zero. As a result, the concept graph forms a sparse structure that preserves only meaningful semantic relationships. Self-loops reflect each concept’s self-similarity, with their strength indicating the internal consistency of the corresponding concept embedding.

### 4.2.2 Graph Modelling

Our Concept Graph model is designed to learn semantically structured representations of images by encoding vision embeddings into disentangled concept nodes. While semantic information is already incorporated into the corresponding concepts at the initial graph level, we apply a series of graph layers to capture the underlying relationships between concepts. This further enables projecting the high-dimensional concept embeddings into a lower-dimensional, more interpretable space, while explicitly modelling inter-concept interactions. The node embeddings of the final graph layer, which preserve the original graph structure but with reduced dimensionality, are then aggregated and passed through a projection layer to produce the final logits for prediction.

In essence, the core idea of our architecture follows the encoder-bottleneck paradigm, where rich visual information is progressively distilled into compact, semantically meaningful graph representations.

#### Architecture

The core of the proposed framework employs a stack of GAT layers to model interactions between concept nodes in the constructed graph. Specifically, the network consists of two sequential GAT layers, each designed to progressively learn the node embeddings while reducing their dimensionality.

This shallow GAT architecture is motivated by empirical findings in prior works [110, 79], which demonstrate that stacking a small number of GAT layers effectively captures local and moderately higher-order dependencies without introducing over-smoothing or overfitting. While deeper GNN architectures theoretically allow information aggregation from more distant nodes, they often suffer from over-smoothing, where node representations become increasingly similar and lose discriminative power. Moreover, excessive depth increases computational complexity and may hinder model convergence, particularly in scenarios with limited data.

Given the limited number of visual concepts in our setting, the resulting graph structure is inherently shallow, consisting of only 10 to 15 nodes. As a result, aggregating information from distant nodes is unnecessary. For robust learning, we adopt a simple yet effective two-layer GAT structure. Each GAT layer is followed by Graph Norm and a non-linear activation function. To ensure stable and balanced learning, Graph Norm is applied prior to graph-level pooling, preventing nodes with excessively large activations from disproportionately influencing the aggregated representation.

Attention coefficient at the  $k$ -th layer and for the  $i$ -th attention head is computed as follows:



$$\alpha_{vu}^{(k,i)} = \frac{\exp\left(\text{LeakyReLU}\left(a^\top \left[\mathbf{W}^{(k,i)}\mathbf{h}_v^{(k-1)} \parallel \mathbf{W}^{(k,i)}\mathbf{h}_u^{(k-1)}\right]\right)\right)}{\sum_{w \in \mathcal{N}(v)} \exp\left(\text{LeakyReLU}\left(a^\top \left[\mathbf{W}^{(k,i)}\mathbf{h}_v^{(k-1)} \parallel \mathbf{W}^{(k,i)}\mathbf{h}_w^{(k-1)}\right]\right)\right)} \quad (4.8)$$

where  $\mathbf{W}^{(k,i)}$  is the learnable weight matrix for the  $i$ -th attention head at layer  $k$ ,  $a$  is a learnable attention vector,  $\parallel$  denotes vector concatenation, and  $\mathcal{N}(v)$  denotes the set of neighbouring nodes of node  $v$ , including  $v$  itself to account for self-loops. This formulation allows the model to compute attention scores based on pairwise feature similarity between nodes, while incorporating learnable parameters and edge weights.

After computing the initial node features  $\mathbf{h}_v^{(0)} \in \mathbb{R}^p$  using Equation 4.5 and the edge weights  $(\mathbf{E}_w)_{vu}$  with Equation 4.7, we can formulate our GAT network as follows.

The first GAT layer with  $n_h$  attention heads is defined as:

$$\tilde{\mathbf{h}}_v^{(1)} = \parallel_{i=1}^{n_h} \sigma \left( \sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(1,i)} \cdot (\mathbf{E}_w)_{vu} \cdot \mathbf{W}^{(1,i)}\mathbf{h}_u^{(0)} \right) \quad (4.9)$$

here  $\mathbf{W}^{(1,i)} \in \mathbb{R}^{p \times m}$  projects the input node features to a lower-dimensional space  $\mathbb{R}^m$ , with  $m < p$ , and  $\sigma(\cdot)$  is the non-linear activation function (ELU). The outputs from both attention heads are concatenated along the feature dimension.

We apply Graph Norm and non-linearity to obtain the updated node features:

$$\mathbf{h}_v^{(1)} = \sigma \left( \text{GraphNorm} \left( \tilde{\mathbf{h}}_v^{(1)} \right) \right) \quad (4.10)$$

The second GAT layer also uses  $n_h$  attention heads, but aggregates them via averaging rather than concatenation:

$$\tilde{\mathbf{h}}_v^{(2)} = \frac{1}{n_h} \sum_{i=1}^{n_h} \sigma \left( \sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(2,i)} \cdot (\mathbf{E}_w)_{vu} \cdot \mathbf{W}^{(2,i)}\mathbf{h}_u^{(1)} \right) \quad (4.11)$$

where  $\mathbf{W}^{(2,i)} \in \mathbb{R}^{m \times z}$  further reduces the feature dimension to  $\mathbb{R}^z$ , with  $z < m$ .

Graph Norm and non-linearity are again applied to obtain the final node representations:

$$\mathbf{h}_v^{(2)} = \sigma \left( \text{GraphNorm} \left( \tilde{\mathbf{h}}_v^{(2)} \right) \right) \quad (4.12)$$

### Projection Layer

After updating all concept nodes  $\mathbf{h}_v^{(2)} \in \mathbb{R}^z$  for each  $v \in \mathcal{V}$  at the final graph layer, the node embeddings are stacked to form the matrix:

$$\mathbf{H}^{(2)} = \begin{bmatrix} \mathbf{h}_1^{(2)} \\ \mathbf{h}_2^{(2)} \\ \vdots \\ \mathbf{h}_r^{(2)} \end{bmatrix} \in \mathbb{R}^{r \times z}$$

where  $r$  is the number of concept nodes and  $z$  is the final node feature dimension.

Next, we apply average pooling over all nodes to obtain a single graph-level embedding:

$$\tilde{\mathbf{h}} = \frac{1}{r} \sum_{v \in \mathcal{V}} \mathbf{h}_v^{(2)} \in \mathbb{R}^z \quad (4.13)$$

The graph-level feature vector  $\tilde{\mathbf{h}}$  is processed by a final projection layer and a softmax function to produce the class probabilities:

$$\hat{\mathbf{y}} = \text{Softmax} \left( \mathbf{W}_{\text{proj}} \tilde{\mathbf{h}} + \mathbf{b}_{\text{proj}} \right) \quad (4.14)$$

where,  $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{out \times z}$  is the learnable weight matrix of the projection layer,  $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{out}$  is the bias term,  $out$  is the number of output classes for the prediction task.

### 4.2.3 Final Training Objective

The final prediction logits are obtained by passing the graph-level embedding through a fully connected projection layer, as described in the previous section. Since our task is formulated as a classification problem, we train the entire framework in an end-to-end manner using the standard categorical cross-entropy for multi-class classification.

Given a batch of  $N$  training examples, let  $y_{i,c} \in \{0, 1\}$  be the ground-truth indicator for class  $c$  of the  $i$ -th example, and let  $\hat{y}_{i,c} \in (0, 1)$  be the predicted probability for class  $c$ , produced by the softmax layer. The categorical cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{out} y_{i,c} \log(\hat{y}_{i,c}) \quad (4.15)$$

To promote sparsity and enhance interpretability, we additionally apply an  $\ell_1$ -regularisation term on the projection layer weights  $\mathbf{W}_{\text{proj}}$ . The final loss function becomes:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \|\mathbf{W}_{\text{proj}}\|_1 \quad (4.16)$$

here  $\lambda$  is a regularization hyperparameter controlling the penalty strength.

The objective of training is to minimise  $\mathcal{L}$  with respect to the learnable parameters of the model, which include the concept graph construction, GAT layers, and final projection layer. The backbone vision model used to extract intermediate activations remains frozen throughout training. Consequently, the learned concept basis, obtained via non-negative matrix factorisation (NMF), is also unaffected by the optimisation process. This design ensures that concept generation remains decoupled from the downstream classification task, promoting stability and interpretability of the learned visual concepts.

## 4.3 Explaining Concept Graphs

Taking advantage of the proposed design structure, our framework provides visually grounded, concept-based explanations at both local and global levels. Once the concept graph for a given image has been constructed and processed by the GAT, we obtain the

final node embeddings, which capture both the semantic content of each visual concept and their interactions. These embeddings are then used to systematically quantify the contribution of each concept and image patch to the model’s output, enabling a structured analysis of the underlying decision-making process.

Overall, our proposed explainability method offers the following properties :

- **Concept-based explanations:** Model decisions are grounded in high-level, interpretable visual concepts, providing global guidance for the model’s decision-making process.
- **Patch-level localisation:** The propagation of concept importance to image patches enables the visualisation of regions that contribute most to the prediction, offering a more interpretable alternative to low-level pixel importance maps.

### 4.3.1 Concept-Level Explanation

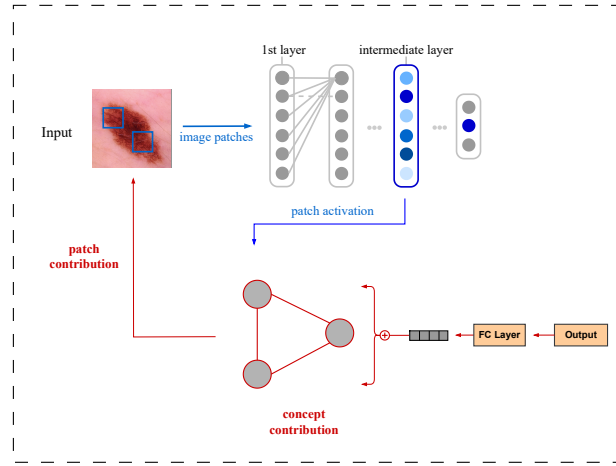


Figure 4.4: Simple diagram of the explanation framework. Nodes in the final layer act as bottlenecks, and the contribution of each patch is traced back from these nodes to the patch-level activations.

After processing the concept graph with the GAT, each concept node  $i \in \mathcal{K}$  is associated with a final node embedding  $\mathbf{h}_i^{(2)} \in \mathbb{R}^z$ , where  $z$  is the reduced embedding dimension. The model’s final prediction is obtained by applying a linear projection to the graph-level representation, as described in the previous section. To compute the importance of each concept for the given class of interest  $c$ , we use the learned projection weights  $\mathbf{W}_{\text{proj}}^{(c)} \in \mathbb{R}^z$ , resulting in the following contribution score for each concept:

$$\text{ConceptScore}(i, c) = \mathbf{W}_{\text{proj}}^{(c)} \cdot \mathbf{h}_i^{(2)} \quad (4.17)$$

This formulation provides a direct and interpretable measure of how much each concept contributes to increasing the logit score for the given class  $c$ . Higher attribution values indicate concepts are more important for the model’s decision-making.

### 4.3.2 Patch-Level Explanation via Concept Propagation

In addition to concept-level explanations, our framework provides patch-level attributions by propagating concept influence down to the image patches. Recall that during the patch pooling step (Equation 4.5), each patch  $j$  is associated with a set of learned concept composition coefficients  $s_{jv}$ , reflecting the degree to which patch  $j$  expresses concept  $v$ . Leveraging the shallow nature of our GAT network, we can compute the overall influence of each patch on the final prediction by applying the chain rule to take the derivative of the prediction with respect to the patch  $j$ . Due to the linear aggregation and projection in our model, the gradient of the class logit with respect to any concept node embedding is constant across nodes. We therefore replace it with a forward-pass-based contribution at the top of the chain rule to better capture each node’s actual influence on the prediction. Doing this, we compute the gradient flow from the prediction back to all final concept nodes, and subsequently from all the concept nodes to the selected patch  $j$ , as informed by the patch pooling mechanism. The limited depth of the model helps prevent gradient saturation, ensuring that the gradient signal remains informative throughout this process. The resulting attribution reflects how the final contribution of each patch depends on both its assigned concept composition coefficients and the downstream sensitivity of the output to the concept nodes. The patch-level attribution can be computed as follows:

$$\text{PatchScore}(j, c) = \sum_{v \in \mathcal{K}} \left( \text{ConceptScore}(v, c) \cdot \frac{\partial \mathbf{h}_v}{\partial g(X_j)} \right) \quad (4.18)$$

This approach enables attribution of model predictions to specific image regions, while preserving alignment with the concept graph structure.

## 4.4 Design Considerations

Determining the size  $r$  of the concept set  $\mathcal{K}$  is a crucial design decision, as it directly influences the expressiveness, interpretability, and class-discriminative power of the resulting concept basis. To guide this choice, we propose a simple unsupervised strategy that favours settings where concepts are both semantically meaningful and class-specific.

More specifically, we aim to identify a value of  $r$  that produces a set of concepts which are (i) *highly discriminative*—that is, each concept is strongly associated with a single class—and (ii) *balanced*—meaning that all classes are represented roughly equally in the set of discriminative concepts. We call the resulting metric the *disentanglement score*, which we use to select the optimal number of concepts  $r^*$ .

The *disentanglement score* is calculated through the multiple steps. We first quantify how class-specific each concept is. Let  $\mathcal{T}_i \subseteq \{1, \dots, N\}$  denote the top- $q\%$  of patches most strongly activating concept  $i$ :

$$\mathcal{T}_i = \{j \mid U_{ij} \geq \tau_i\}, \quad \tau_i = P_{100(1-q)}(U_i). \quad (4.19)$$

Given *out* classes, we compute the class distribution of these top-activating patches:

$$R_{ic} = \frac{1}{|\mathcal{T}_i|} \sum_{j \in \mathcal{T}_i} \mathbf{1}[y_j = c], \quad c \in \{1, \dots, \text{out}\}.$$

The *discriminativity* of concept  $i$  is defined as the dominance of its most frequent class:

$$D_i = \max_c R_{ic}, \quad (4.20)$$

where  $D_i \in [\frac{1}{out}, 1]$ . Concepts with  $D_i$  close to  $\frac{1}{out}$  are considered *class-agnostic*.

We assign each concept to its dominant class  $c_i^* = \arg \max_c R_{ic}$  if  $D_i$  exceeds a predefined threshold  $\theta$ ; otherwise, it is treated as ambiguous and excluded from further analysis.

Let  $\mathcal{D} = \{i \mid D_i \geq \theta\}$  denote the set of *discriminative* concepts, with size  $d = |\mathcal{D}|$ . For each class  $c$ , let  $n_c = |\{i \in \mathcal{D} \mid c_i^* = c\}|$  be the number of discriminative concepts assigned to that class. We define the ideal per-class allocation as  $\bar{n} = d/out$ , and propose the following *disentanglement score*:

$$D(r) = \frac{1}{d} \sum_{i \in \mathcal{D}} D_i - \lambda \cdot \frac{1}{out} \sum_{c=1}^{out} \left| \frac{n_c}{d} - \frac{1}{out} \right|, \quad (4.21)$$

where  $\lambda \geq 0$  controls the strength of the penalty for class imbalance.

The first term encourages highly class-specific concepts, while the second penalizes unbalanced allocations. Class-agnostic concepts (those with  $D_i < \theta$ ) do not contribute to the score. We compute this score for different values of  $r$  and select  $r^* = \arg \max_r D(r)$  as the optimal number of concepts.

When  $D_i \approx 1$  for most concepts in  $\mathcal{D}$ , and  $n_c \approx \bar{n}$  for all classes, the score is maximized—indicating a highly disentangled and well-balanced concept set. Conversely, if many concepts are ambiguous or heavily skewed toward a few classes, the score is lower, suggesting that a different  $r$  would yield better interpretability.

In practice, we observe that for top-10% and  $\lambda = 1.0$ , class-aligned concept structures tend to emerge when  $r \in [5, 15]$ , yielding the best trade-off between semantic clarity and class discrimination. Notably, this selection method prioritizes interpretability and class alignment over predictive performance.

## 4.5 Further Evaluation

### Patch Localisation Evaluation

To evaluate the spatial fidelity of patch-level localization produced by our model, we quantify the alignment between conceptually important patches and pathologically relevant areas. The relevant regions are defined by segmentation masks generated using DeepLabV3 [13], a state-of-the-art semantic segmentation model trained on the *HAM10000* dataset [106]. More precisely, we aim to understand what proportion of the top-ranked patches (as identified by our model) fall within the lesion area highlighted by the segmentation.

Let  $S \subset \mathcal{P}$  denote the set of patch indices whose spatial regions intersect with the segmentation mask, and let  $T \subset \mathcal{P}$  be the set of top- $|S|$  most important patches as ranked by the model. To ensure meaningful alignment, we introduce an overlap threshold  $\alpha \in [0, 1]$ . A top patch is considered valid only if at least  $\alpha$  proportion of its area overlaps with the segmentation mask.

Formally, for each patch  $p \in T$ , let  $A(p)$  be the area of the patch and  $A(p \cap \text{Seg})$  the area of its intersection with the segmentation mask. We define the set of valid patches as:

$$T_\alpha = \left\{ p \in T \mid \frac{A(p \cap \mathbf{Seg})}{A(p)} \geq \alpha \right\}$$

The *localisation ratio*  $R_\alpha$  is then defined as:

$$R_\alpha = \frac{|S \cap T_\alpha|}{|S|}$$

This metric captures the extent to which top-ranked patches align with clinically meaningful lesion regions, providing insight into the correspondence between model explanations and medical relevance. By varying the overlap threshold  $\alpha$ , we can further explore the trade-off between explanation sparsity and localisation fidelity.

---

# Chapter 5

## Experiments and Results

In the recent years, deep neural networks with concept bottlenecks have re-emerged as an essential line of research for building inherently explainable models. While effective, this approach relies on expert-annotated concepts for each image, which limits scalability. Moreover, it is based on a critical assumption: the pre-trained model already implicitly encodes the predefined textual concepts (i.e., clinical annotations). Consequently, the faithfulness of existing concept predictors to their underlying concepts remains a challenge. This, in turn, can make concept bottleneck interpretability fragile, as it may occasionally depend on spurious semantic features.

To address these, we propose building an interpretable model based solely on visual features. In this chapter, we present our empirical findings on constructing a part-based concept bottleneck model with a graph neural network, specifically in the context of medical image analysis. Through a series of quantitative and qualitative experiments, we evaluate the effectiveness of the proposed framework.

### 5.1 Dataset

To evaluate our framework in both general-purpose and domain-specific settings, we employ four publicly available datasets. These datasets span natural image classification and medical imaging tasks, with a particular focus on skin lesion analysis.

- **ImageNet** [58]; A large-scale image dataset widely used for benchmarking deep learning models in image classification. It contains approximately 1.2 million labeled images across 1,000 object categories.
- **HAM10000** [106]; A comprehensive dermoscopic image dataset targeting pigmented skin lesions with 7 categories. It consists of over 10,000 high-resolution images with expert annotations, supporting research on automated skin lesion detection and classification.
- **Derm7pt** [89]; A curated dataset of 1,011 dermoscopic images annotated with

clinical metadata based on seven dermoscopic concepts. It serves as a valuable resource for interpretable skin lesion classification research.

- **PH<sup>2</sup>** [75]; A dermatology research dataset containing 200 dermoscopic images for melanoma detection. The dataset provides detailed diagnostic labels and lesion segmentation masks.

The *Derm7pt* and *PH<sup>2</sup>* datasets serve as the main benchmarks in our experiments. Following Patrício et al. [83], we partition both datasets into two classes: *Nevus* and *Melanoma*. The *Derm7pt* dataset is filtered to obtain 827 dermoscopic images, which are then divided into the *Nevus* and *Melanoma* classes. The *PH<sup>2</sup>* dataset consists of 200 dermoscopic images, including 40 *Melanomas*, 80 *Common Nevi*, and 80 *Atypical Nevi*. As in previous works, both *Common* and *Atypical Nevi* are grouped into a single *Nevus* class.

We also include the *HAM10000* dataset in our experiments. In line with previous work [7, 36], we restrict our analysis to two dermoscopic image types for binary classification: approximately 1.1k *Melanoma* images representing the malignant class, and approximately 6k *Melanocytic Nevi* images representing the benign class.

### 5.1.1 Data Preparation

In our experiments, each dataset is partitioned into 70% training, 15% validation, and 15% test subsets. However, since the *PH<sup>2</sup>*, *Derm7pt*, and *HAM10000* datasets exhibit imbalanced class distributions, we trained on a balanced version of the data, while the test and validation sets remained in their original state. To achieve this, we oversampled the under-represented classes by including additional copies of their samples. Importantly, these were not exact duplicates: each copy was augmented with transformations from the training pipeline. Moreover, to achieve a fair concept generation for medical datasets, NMF was also applied to a balanced dataset split.

For the medical datasets (*PH<sup>2</sup>*, *Derm7pt*, *HAM10000*), the training augmentation consisted of resizing, center cropping, random 90° rotations, and horizontal flips, while evaluation used only resizing and center cropping. For the *ImageNet* dataset, we applied the standard augmentation scheme of random resized cropping, horizontal flipping, and color jitter during training, and resizing with center cropping for evaluation. No additional augmentation is applied to the validation and test splits in any of the datasets.

## 5.2 Experiment Setting

### 5.2.1 Backbone Models

Our framework consists of two main stages, as described in Chapter 4: *Concept Basis Creation* and *Learning with Concept Graphs*. In the first stage, we generate the concept basis, and in the second stage, we represent images as concept graphs based on the learned concept basis and perform end-to-end training. One of the most essential components of both stages is the backbone model, through which we obtain the latent visual embeddings at an intermediate layer, denoted as  $g(\cdot)$ . This backbone model remains frozen during training and is only used to extract visual cues.

While the backbone models we employ are deep vision architectures capable of extracting visual patterns across different datasets, we experiment with multiple backbone models



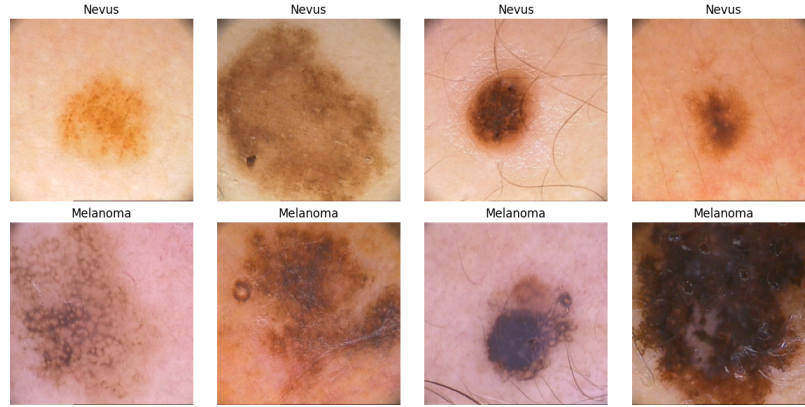


Figure 5.1: Example dermoscopic images from the  $PH^2$  dataset. The top row shows *Nevus* samples, while the bottom row displays *Melanoma* images.

to investigate how the quality of concept generation is influenced. For this purpose, we use the DenseNet-201 [46], ResNet50 [40], and MobileNet-V2 [92] architectures. ResNet50 serves as our primary backbone model for comparative purposes, as it is the best-performing model and has been widely used in related studies. The remaining models are included to assess the impact of the backbone choice on overall performance.

- **ResNet-50;** Is a widely used deep CNN model consisting of 50 layers. It relies on residual connections, which allow the network to bypass certain layers during training and help address the vanishing gradient problem. Widely regarded for its effectiveness in image classification tasks, it serves as a reliable baseline in my research studies.
- **DenseNet-201;** Is a deep CNN architecture with 201 layers. It uses dense connections, where each layer takes in feature maps from previous layers, facilitating feature reuse and improving gradient propagation. While capable of extracting detailed features, it can be memory intensive.
- **MobileNet-V2;** While it has a similar range of depth as the ResNet-50 model, it relies on inverted residuals and linear bottlenecks to reduce computational overhead. While it sacrifices some accuracy compared to larger models, it offers significant advantages in speed and memory usage.

All backbone models are pre-trained on the ImageNet dataset to ensure that they are capable of extracting general visual features relevant to downstream tasks. For DenseNet-201 and MobileNet-V2, we use the output of the final convolutional layer as the intermediate feature representation. For ResNet-50, we use the output of the last convolutional block, prior to global average pooling. The pre-training enables the backbones to provide meaningful latent representations, which serve as the foundation for both concept extraction and subsequent concept graph construction in our framework.

### 5.2.2 Concept Number and Patch Size

Determining the appropriate patch size and the number of concepts  $r$  is a crucial design choice, as it directly affects the fidelity, interpretability, and discriminative power of the

learned concept basis. To investigate the effect of the number of concepts and patch size on prediction performance, we conducted a series of experiments on the  $PH^2$  dataset. In these experiments, we systematically varied both the patch size and the number of concepts and evaluated their impact on both model performance and the quality of the extracted concepts.

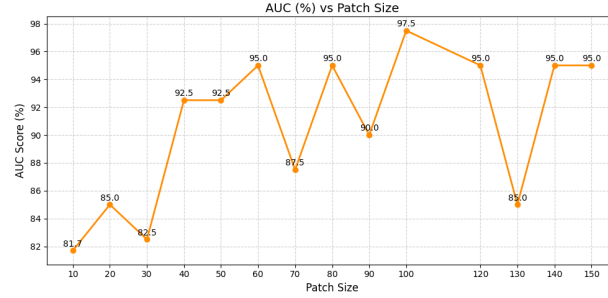


Figure 5.2: Shows the AUC scores of our proposed model under varying numbers of patch sizes used for concept generation and formulation. The concept number in this experiment is fixed to 15.

In the first experiment, we assessed the effect of patch size by varying its value while keeping the number of concepts fixed at 15. We observed the model performance (measured in AUC score) under varying patch sizes from  $10 \times 10$  to  $150 \times 150$ . Our results indicate that while slightly larger patches tend to yield better results, the performance remains relatively stable across a range of patch sizes, as illustrated in Figure 5.2.

In a second experiment, we fixed the patch size at  $70 \times 70$  and varied the number of extracted concepts from 2 to 40. As can be seen in Figure 5.3, the AUC score remains stable across different concept counts, indicating that the model’s predictive performance is not highly sensitive to the number of concepts. For further evaluation, we evaluated the quality of concept generation (Figure 5.4). A low number of concepts (e.g., fewer than 10) results in general but non-specific concepts, which limits interpretability. Increasing the number of concepts improves concept reconstruction quality and sparsity, but when the number exceeds 20, the representations tend to become less stable, likely due to over-fragmentation or redundancy. This reflects a typical behaviour in NMF, where too many components may dilute the semantic coherence of the concept basis.

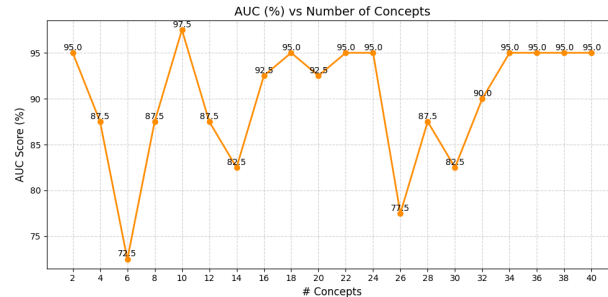


Figure 5.3: Shows the AUC scores of our proposed model under varying numbers of concepts used for concept generation and formulation. The patch size in this experiment is fixed at  $70 \times 70$ .

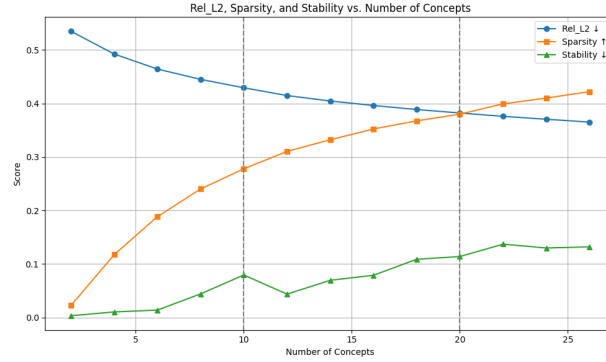


Figure 5.4: Shows the reconstruction error, sparsity, and stability scores of our proposed model under varying numbers of concepts used for concept generation and formulation. The patch size in this experiment is fixed at  $70 \times 70$ .

Our initial analysis suggested that our cropping and resizing procedure  $\pi(\cdot)$  is robust, and the overall model performance does not necessarily depend on precise patch granularity. Therefore, patch sizes can be selected flexibly within a reasonable range, balancing local detail and semantic abstraction. To that end, we set the patch size to  $70 \times 70$  in all our experiments. However, it is also worth noting that this flexibility may be attributed to the nature of medical dermatology datasets, where semantic cues are often localized and concentrated near the centre of the image.

In contrast, the design consideration for selecting the number of concepts is not as straightforward for several reasons: a larger number of concepts dilutes semantic coherence; the choice of concept number should not be driven by predictive performance but by explainability; and the number of concepts should ideally be selected automatically. Therefore, to determine the number of concepts  $r$ , we followed the unsupervised selection strategy described in Section 4.4. This approach aims to identify the most disentangled set of concepts, thereby enhancing both the expressiveness and semantic clarity of the generated concept basis.

For example, for the  $PH^2$  dataset, using 10 concepts yielded the highest *disentanglement score*, as shown in Table 5.1. This result reflects an optimal trade-off between concept discriminativity and class balance for the binary classification case (*Nevus* and *Melanoma*). Specifically, the model achieved a balanced class distribution of  $[4, 4]$  among the top concepts, with only two concepts identified as class-agnostic or associated with more generic features. This suggests that the discriminative concepts are well distributed across both classes. Moreover, the average per-concept discriminativity ( $\text{Avg } D_i = 0.7413$ ) is relatively high, meaning that most concepts are strongly aligned with a single class. See Table 5.1 for the full comparison.

Nevertheless, it is important to note that a high average discriminativity score does not imply complete disentanglement. Due to the additive nature of non-negative matrix factorization, some degree of overlap or semantic blending may persist—especially for visual patterns shared across different lesion types. Our objective is to select the most disentangled concept representations possible to ensure coherent and interpretable concept generation.

# Concepts	Avg $D_i$	Class Split	Penalty	Score
6	0.7373	[4, 2]	0.1667	0.5706
7	0.7471	[4, 2]	0.1667	0.5804
8	0.7951	[3, 2]	0.1000	0.6951
9	0.7361	[4, 4]	0.0000	<u>0.7361</u>
10	0.7413	[4, 4]	0.0000	<b>0.7413</b>
11	0.7649	[3, 4]	0.0714	0.6935
12	0.7531	[4, 5]	0.0556	0.6975

Table 5.1: Disentanglement score statistics on PH2 dataset for concept counts from 6 to 12. Best score is given in **bold**, second-best score is underlined. We set the  $\lambda = 1.0$  and top-10% in our experiment.

### 5.3 Quantitative Evaluation

The goal of this section is to study the effectiveness of our proposed framework for interpretability. To that end, we examine the following questions:

1. Can concept extraction methods be effectively employed for concept generation in a medical imaging setting? If so, how effective is the method we utilize compared to others?
2. At what cost does interpretability come? Is the drop in prediction accuracy negligible?
3. How does our method compare with other concept-based explainability approaches in the medical domain?
4. Analysis of Explainability and Localisation: Are the learned concepts and localisation faithful to the prediction process?

#### 5.3.1 Evaluation of Concept Generation

While concept extraction methods have been successfully applied to general-purpose datasets such as *ImageNet*, the quality of concept generation in these settings has not often been scrutinised. This is largely because concept quality can be easily assessed in such datasets due to the rich, high-level semantic information present in natural images.

In our framework, concept extraction serves as a fundamental pillar. However, categorising and assessing skin lesion datasets can be considerably more challenging compared to general-purpose datasets, as skin lesion images often exhibit more fine-grained and subtle visual details. Therefore, it is critical to ensure that concept extraction techniques used for concept generation are indeed applicable to medical imaging datasets in the first place.

Before proceeding with our method, we first aim to evaluate the extent to which established concept extraction techniques—namely PCA, K-Means, and NMF—can be applied to medical imaging datasets, specifically skin lesion datasets. To this end, we compare the performance of these concept extraction methods on both a semantically rich dataset (*ImageNet*) and a skin lesion dataset. We employ a set of quantitative metrics, including *sparsity*, *reconstruction error*, *stability*, *FID*, and *out-of-distribution (OOD)* scores

(as discussed in Section 3.3), to assess the effectiveness of each method in both general and medical imaging contexts.

	Relative $\ell_2$ ( $\downarrow$ )	Sparsity ( $\uparrow$ )	Stability ( $\downarrow$ )	FID ( $\downarrow$ )	OOD ( $\downarrow$ )
	Den / R50 / Mob	Den / R50 / Mob	Den / R50 / Mob	Den / R50 / Mob	Den / R50 / Mob
<b>PCA</b>	0.59 / 0.38 / 0.43	0.00 / 0.00 / 0.00	0.33 / 0.28 / 0.36	47.7 / 11.6 / 12.1	0.34 / 0.15 / 0.19
<b>K-Means</b>	0.66 / 0.50 / 0.54	0.96 / 0.96 / 0.96	0.04 / 0.01 / 0.03	94.0 / 43.9 / 41.3	0.28 / 0.14 / 0.17
<b>NMF</b>	0.58 / 0.40 / 0.46	0.41 / 0.36 / 0.41	0.14 / 0.09 / 0.12	69.2 / 23.8 / 25.4	0.32 / 0.16 / 0.20

Table 5.2: Concept extraction comparison on *ImageNet*. Den, R50, and Mob denote DenseNet-201, ResNet-50, and MobileNet-V2. Concepts are extracted from the final activation layer of the networks. Results are obtained from a set of  $\sim 2.5$ k images for two semantically similar classes, *Ambulance* and *Recreational Vehicle*.

Across both datasets, the behaviour and relative performance of PCA, K-Means, and NMF remain consistent (see Tables 5.2 and 5.3). We also observed that ResNet-50 consistently outperforms the other backbone models in terms of relative error, sparsity, stability, and out-of-distribution (OOD) scores across both datasets. In summary, our findings demonstrate that while PCA consistently achieves the best reconstruction accuracy, it produces dense, less interpretable representations and lacks stability. K-Means maintains very sparse, highly interpretable concept structures, though it continues to struggle with reconstruction error and plausibility. NMF consistently offers a balanced compromise, combining moderate sparsity with strong reconstruction quality and stable, interpretable concepts. This, therefore, justifies our choice of NMF for concept basis generation.

	Relative $\ell_2$ ( $\downarrow$ )	Sparsity ( $\uparrow$ )	Stability ( $\downarrow$ )	FID ( $\downarrow$ )	OOD ( $\downarrow$ )
	Den / R50 / Mob	Den / R50 / Mob	Den / R50 / Mob	Den / R50 / Mob	Den / R50 / Mob
<b>PCA</b>	0.41 / 0.27 / 0.30	0.00 / 0.00 / 0.00	0.33 / 0.37 / 0.39	7.36 / 1.44 / 1.18	0.13 / 0.06 / 0.07
<b>K-Means</b>	0.50 / 0.37 / 0.41	0.96 / 0.96 / 0.96	0.04 / 0.02 / 0.03	21.1 / 7.38 / 5.15	0.10 / 0.05 / 0.05
<b>NMF</b>	0.40 / 0.28 / 0.32	0.34 / 0.35 / 0.40	0.09 / 0.07 / 0.11	13.4 / 3.21 / 2.87	0.11 / 0.06 / 0.07

Table 5.3: Concept extraction comparison on the large-scale dermoscopic image dataset of pigmented lesions, *HAM10000*. Den, R50, and Mob denote DenseNet201, ResNet50, and MobileNetV2. Concepts are extracted from the final activation layer of the networks. Results are obtained from a set of  $\sim 10$ k images from all classes of *HAM10000*.

Notably, both reconstruction error and FID scores are generally lower on the *HAM10000* dataset compared to *ImageNet*, likely due to the simpler and more homogeneous nature of medical images. More importantly, our findings demonstrate that concept extraction methods originally developed for natural image datasets like *ImageNet* can be successfully adapted to medical imaging tasks such as those in *HAM10000*.

Overall, this suggests that concept-based explanations using NMF are transferable and effective in medical contexts, providing a strong foundation for interpretable, concept-driven AI in healthcare applications.

### 5.3.2 Performance Trade-off

The advantage of our model over black-box methods lies in its inherent interpretability. By using unsupervised visual concepts to build a structured, patch-based graph representation, our model offers an insight into how and why a decision is made—something the standard ResNet model lacks. However, in this section, we aim to examine the cost

at which this interpretability comes. Is the drop in prediction accuracy negligible? To answer this, we conduct experiments on both the *ImageNet* and *HAM10000* datasets.

The results across both datasets reveal differing scales of performance trade-off between the standard classification model (baseline ResNet-50) and our proposed part-based Concept-Graph bottleneck model.

On the *ImageNet* subset (Table 5.4), which includes semantically similar classes (*Ambulance* and *Recreational Vehicle*), the ResNet-50 baseline significantly outperforms all concept-based variants across all metrics—AUC, Accuracy, and F1—achieving around 97.0% respectively. In contrast, all versions of our model exhibit a consistent drop around 1.0-3.0% in performance across all metrics, while still showing significant prediction ability, around  $\sim 96.0\%$  accuracy. Among our backbones, DenseNet slightly outperforms the others, although it still lags behind the baseline. MobileNetV2 shows the lowest accuracy, which may be attributed to its lower representational capacity and higher variance, as reflected in its broader standard deviation margins.

	AUC	ACC	F1
R50 (baseline)	<u>97.58</u> <sub>0.35</sub>	<u>97.44</u> <sub>0.32</sub>	<u>97.36</u> <sub>0.44</sub>
Ours (R50)	96.37 <sub>0.54</sub>	94.55 <sub>0.24</sub>	95.55 <sub>0.42</sub>
Ours (DenseNet)	96.90 <sub>0.68</sub>	95.34 <sub>0.67</sub>	95.81 <sub>0.71</sub>
Ours (MobileNet)	96.10 <sub>0.95</sub>	93.56 <sub>0.62</sub>	94.70 <sub>1.02</sub>

Table 5.4: Performance comparison between ResNet-50 and our proposed part-based Concept-Graph bottleneck model using three different backbone architectures on two semantically similar classes (*Ambulance* and *Recreational Vehicle*) of ImageNet. Underline indicates the best result. The performance is reported as mean<sub>std</sub> across multiple runs on test set.

The *HAM10000* results (Table 5.5) show a similar outcome. While the ResNet-50 baseline again achieves the highest AUC (93.57%), ACC (92.35%) and F1 score (90.34%), our concept-based model delivers comparable results, without significant performance drop. This suggests that the Concept-Graph model is capable of achieving acceptable prediction accuracy on medical datasets as well. Among the backbones, the DenseNet201-based model offers the best trade-off, with ResNet-50 following closely. Similarly, MobileNetV2 underperforms slightly across all metrics, though the gap is narrow.

The performance gap observed between baseline model and our model in both datasets might highlight the limitations of visual concept bottlenecks in handling complex, large-scale tasks. Two key factors likely contribute to the performance gap. First, concept extraction quality can be lower due to high visual diversity and less structured semantics. Second, our patch-based graph architecture may struggle to capture long-range dependencies and spatial relationships when semantic information is scattered across the image. Moreover, the fragmentation of input images into patches can lead to a loss of important spatial context. These findings suggest that while our method is suitable for medical imaging tasks with localized semantics, further architectural improvements are needed to make it effective on more complex datasets.

### 5.3.3 Benchmark Comparison

This section presents a quantitative comparison between our method and existing concept-based explainability approaches on two skin lesion diagnosis datasets: *Derm7pt*

	AUC	ACC	F1
R50 (baseline)	<u>93.57</u> <sub>0.95</sub>	<u>92.35</u> <sub>0.82</sub>	<u>90.34</u> <sub>0.74</sub>
Ours (R50)	91.23 <sub>0.94</sub>	87.03 <sub>0.55</sub>	87.89 <sub>0.40</sub>
Ours (DenseNet)	91.83 <sub>0.54</sub>	88.67 <sub>0.79</sub>	88.45 <sub>0.49</sub>
Ours (MobileNet)	90.64 <sub>1.17</sub>	86.01 <sub>1.08</sub>	87.18 <sub>1.34</sub>

Table 5.5: Performance comparison between ResNet-50 and our proposed part-based Concept-Graph bottleneck model using three different backbone architectures on the *HAM10000* skin lesion dataset. Models are trained on the *Melanoma* and *Melanocytic Nevi* classes. Underline indicates the best result. The performance is reported as mean<sub>std</sub> across multiple runs on test set.

and  $PH^2$ . Following prior work, we evaluate all methods under the same binary classification task (*Nevus* vs. *Melanoma*). The results are compared on test set. Unlike previous approaches that rely heavily on supervised concept annotations—such as clinical attributes or manually defined bottlenecks—our framework is entirely unsupervised and visually grounded. The objective of this comparison is to assess how our method performs relative to supervised alternatives (see Table 5.6).

Method	Derm7pt			PH <sup>2</sup>		
	AUC	ACC	F1	AUC	ACC	F1
Sarkar et al. [93]	76.22 <sub>2.06</sub>	73.89 <sub>1.47</sub>	66.81 <sub>1.23</sub>	79.33 <sub>0.62</sub>	88.00 <sub>3.26</sub>	79.66 <sub>2.11</sub>
PCBM [119]	72.96 <sub>1.44</sub>	76.98 <sub>1.21</sub>	71.04 <sub>1.38</sub>	78.33 <sub>1.17</sub>	89.33 <sub>1.89</sub>	81.49 <sub>2.57</sub>
PCBM-h [119]	83.27 <sub>1.14</sub>	79.89 <sub>0.89</sub>	74.48 <sub>1.37</sub>	92.32 <sub>1.47</sub>	90.67 <sub>1.89</sub>	83.30 <sub>2.55</sub>
CBE [83]	76.60 <sub>0.35</sub>	83.75 <sub>0.26</sub>	78.13 <sub>0.44</sub>	97.59 <sub>0.00</sub>	<u>96.00</u> <sub>0.00</sub>	93.89 <sub>0.00</sub>
MICA (w/ bot) [7]	84.11 <sub>1.10</sub>	82.20 <sub>1.31</sub>	78.08 <sub>1.22</sub>	97.66 <sub>1.24</sub>	<u>96.00</u> <sub>3.26</sub>	<u>94.40</u> <sub>1.48</sub>
MICA (w/o bot) [7]	85.59 <sub>1.11</sub>	83.94 <sub>1.31</sub>	79.38 <sub>1.22</sub>	<b>98.18</b> <sub>1.43</sub>	<b>98.67</b> <sub>1.89</sub>	<b>95.34</b> <sub>1.18</sub>
CW [14]	86.50 <sub>0.40</sub>	83.85 <sub>0.48</sub>	80.00 <sub>0.75</sub>	—	—	—
CAW [44]	<b>88.60</b> <sub>0.10</sub>	<b>84.79</b> <sub>0.79</sub>	<b>81.34</b> <sub>0.85</sub>	—	—	—
Ours	84.76 <sub>1.37</sub>	81.48 <sub>0.71</sub>	<u>80.69</u> <sub>0.60</sub>	88.81 <sub>2.14</sub>	87.20 <sub>1.78</sub>	91.60 <sub>1.67</sub>

Table 5.6: Quantitative comparison of our method against state-of-the-art supervised concept-based approaches for skin cancer diagnosis. All methods, including ours, are built upon the ResNet-50 architecture for feature extraction to ensure a fair comparison. Results are reported as mean<sub>std</sub> across multiple runs on the test set. The best score is shown in **bold**, and the second-best is underlined. The benchmark results are reported as stated in the corresponding publications.

To aid the comparison, we briefly outline how each compared method incorporates concepts within the bottleneck modelling paradigm, including their supervision type and use of concept representations.

- Sarkar et al. [93]: An ante-hoc CBM that uses manually annotated clinical concepts as a bottleneck. Concepts are explicitly supervised and directly used for prediction.
- PCBM / PCBM-h [119]: Post-hoc concept bottleneck models that project image embeddings onto a concept subspace defined by CAVs or vision-language embeddings. PCBM uses these projections for prediction, while PCBM-h adds a residual

connection from embedding space to sparse prediction layer to capture information not explained by the concept space.

CBE [83]: A supervised CBM that disentangles CNN feature maps into concept-specific filters using clinical annotations. It enforces uniqueness, semantic alignment, and spatial coherence to ensure that concepts are interpretable and localized. It is tailored for medical diagnosis, relying on manually defined clinical concepts.

- MICA (w/ bottleneck) [7]: A supervised CBM using a multi-level image–concept alignment strategy. It depends on ground-truth concept annotations per image and uses a concept bottleneck structure to make predictions.
- MICA (w/o bottleneck): A weaker bottleneck-free variant of MICA that still aligns image regions, concept tokens, and diagnostic labels across multiple levels, but removes the concept prediction constraint, offering more flexibility. However, it still requires concept annotations for alignment, making it non-scalable to datasets without labelled concepts.
- Concept Whitening (CW) [14]: Not a CBM in the strict sense, but a technique that rotates and decorrelates latent representations so that each feature axis corresponds to a disentangled concept direction. It uses concept supervision to ensure that each whitened direction captures a distinct concept.
- Concept Attention Whitening (CAW) [44]: An extension of CW that applies attention to enhance the alignment between concepts and specific image regions, improving interpretability while preserving classification performance.
- Ours: A fully unsupervised, visually grounded CBM that discovers and learns concepts directly from images without any concept annotations, offering scalable interpretability.

As shown in Table 5.6, the best overall performance on *Derm7pt* is achieved by CAW method, with the highest AUC (88.60), ACC (84.79), and F1 (81.34) scores. Both CW and MICA (w/o bot) methods follow closely, delivering strong results but still relying on annotated or textual concept supervision. While our method does not surpass the top-performing approaches, it remains competitive, achieving AUC: 84.76, ACC: 81.48, and F1: 80.69. This places it on par with or ahead of several earlier methods, including methods such as PCBM, CBE, and MICA (w/ bot). On  $PH^2$  dataset, our model shows reasonable AUC (88.81), ACC (87.20), and F1 (91.60) scores, despite using no expert supervision. This marks an improvement over the performance of CBM-based models, specifically PCBM and PCBM-h methods. Overall, these results highlight our model’s ability to generalize and perform reliably even on small datasets. Our model performs slightly better on the *Derm7pt* dataset compared to the  $PH^2$  dataset. This difference may be due to the smaller size of  $PH^2$ , which makes it less effective for concept learning with NMF.

Although our model falls behind the top performing methods, it performs comparably to top supervised concept bottleneck models, despite requiring no annotated concepts. This competitive performance is particularly notable given its fully unsupervised concept discovery pipeline. However, it is worth noting that, unlike in the other datasets (i.e., *ImageNet* and *HAM1000*), the relatively larger performance gap between our model and the benchmarks may be attributed to the small sample size of the benchmark dataset, which likely limits the model’s ability to learn more robust and discriminative concept



representations. This limitation is potentially offset in other methods by the use of clinical annotation, which may compensate for the lack of a large dataset.

In essence, our approach addresses a core issue of conventional concept bottleneck models: the assumption that pre-trained deep neural networks inherently encode human-defined textual concepts (e.g., clinical annotations). This assumption is often unverified and can lead to incomplete or spurious mappings, undermining the faithfulness of the model’s interpretability. In this sense, our framework embraces the idea of explaining vision models through visual concepts—rather than imposing textual labels—which may offer a more faithful and robust path to interpretability. This is especially important in domains like medical imaging, where semantic alignment between visual evidence and textual concepts is often not validated.

Overall, the effectiveness of our pipeline in visual concept discovery, along with the considerable performance of our overall framework on both datasets, demonstrates that visually grounded concept bottlenecks can offer a viable and interpretable alternative to traditional supervised approaches. With that in mind, our method aims to encourage research toward scalable and trustworthy explainable AI in healthcare, particularly in settings where expert annotations are limited, costly, or inconsistent.

### 5.3.4 Faithfulness Analysis

#### Fidelity Test

To assess whether the identified concepts and localized patches are faithful to the model’s decision-making process, we conduct a series of experiments using the fidelity metrics discussed in Section 3.3, following the prior works [7, 25]. These metrics evaluate the change in the model’s prediction confidence when the most important concepts or spatial regions are either removed or added.

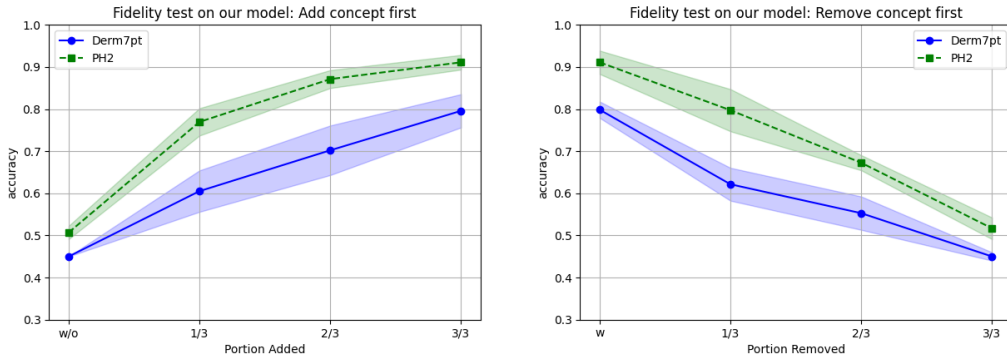


Figure 5.5: Concept-level fidelity analysis of our framework. **Left:** Insertion curve. **Right:** Deletion curve. Results are shown for the *Derm7pt* and *PH<sup>2</sup>* datasets. Shaded regions around the curves indicate standard deviation.

For patch localization, we do not apply pixel-level perturbations. Instead, we skip the contribution of specific patches during the graph pooling phase and observe how this omission affects the final prediction derived from the graph representation. This enables evaluation while remaining robust to perturbation bias [50]. A similar procedure is used for concept nodes: we exclude the most influential concept node from the

graph to observe its impact. Since our GAT model is permutation-invariant, with node embeddings aggregated via averaging, these intervention is used to test the importance of individual patches and concepts.

We first evaluate the faithfulness of our Concept-Graph model on the benchmark datasets *Derm7pt* and *PH<sup>2</sup>*. Our results in Figures 5.5 and 5.6 show that removing the most important concept or spatial patch leads to a noticeable decrease in predictive performance, while adding them significantly improves accuracy. These findings suggest that both localized patches and concept nodes play a meaningful role in model decisions.

### Semantic alignment with baseline CNN

We further test faithfulness on a general-purpose dataset, *ImageNet*, to investigate whether our model’s logic aligns with that of a standard CNN, using fidelity test. Specifically, we analyse whether the important patch locations identified by our Concept-Graph model also correspond to the regions deemed important by a baseline ResNet-50 model. To do so, we perturb the input image to the ResNet model by replacing regions of image identified as important by our model with the average of their surrounding pixels. We observe that removing the top-ranked patches according to our model also reduces the prediction confidence of the baseline CNN. Conversely, preserving or adding these regions improves prediction. This analysis shows that our model focuses on similar semantic regions as a standard CNN, thereby validating its faithfulness to conventional predictive behaviour despite its inherently different structure.

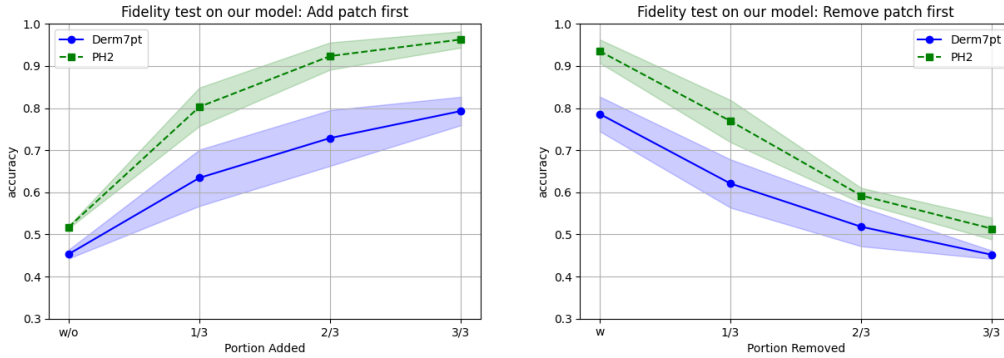


Figure 5.6: Patch-level fidelity analysis of our framework. **Left:** Insertion curve. **Right:** Deletion curve. Results are shown for the *Derm7pt* and *PH<sup>2</sup>* datasets. Shaded regions around the curves indicate standard deviation.

### Further patch-localization evaluation

To quantify the spatial alignment between model explanations and clinically meaningful lesion areas, we applied our localisation ratio metric  $R_\alpha$ , see Section 4.5. This metric measures the proportion of top-ranked patches (as identified by our model) that fall within the lesion region highlighted by segmentation masks. As shown in Table 5.7, we observe a consistent trend across both datasets: the localisation ratio decreases as the overlap threshold  $\alpha$  increases. This reflects a natural trade-off, higher overlap demands stricter spatial correspondence. Notably, both datasets exhibit similar alignment at all threshold values, with the highest alignment at  $\alpha = 0.10$  reaching around 75-80%. Even

at the threshold ratio  $\alpha = 0.50$ , roughly half of the most influential patches still fall within the segmentation region. These results support the spatial faithfulness of our model’s patch-level importance, especially under moderate alignment criteria.

Dataset	$R_{0.10}$	$R_{0.25}$	$R_{0.50}$
Derm7pt	0.7911	0.6728	0.4813
PH2	0.7432	0.6418	0.4722

Table 5.7: Localisation ratio  $R_\alpha$  across different overlap thresholds  $\alpha \in \{0.10, 0.25, 0.50\}$  using patch size 70 and stride 0.5.

Overall, our observations indicate that the regions identified by our Concept-Graph model align well with the semantically meaningful areas utilized by traditional CNNs. Importantly, this is achieved while also providing structured, interpretable representations in the form of concept graphs, highlighting the faithfulness of our approach. However, it is worth noting that our model exhibits higher variance across different runs, suggesting minor stability issues. This variability likely stems from the concept generation process, as reflected in the concept stability scores reported earlier.

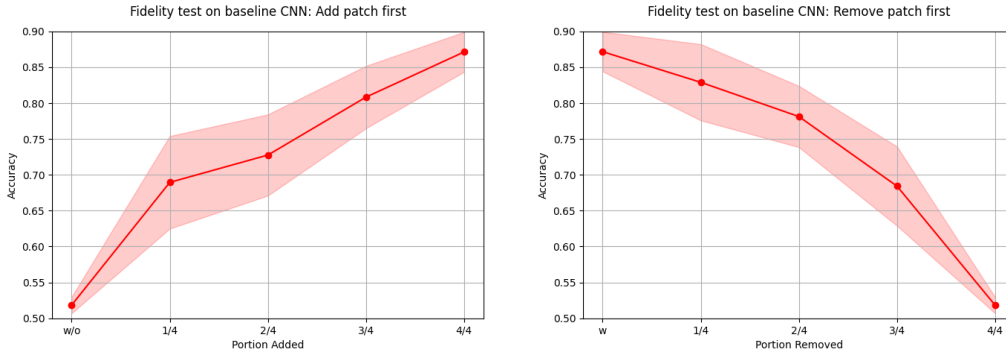


Figure 5.7: Fidelity analysis for semantic alignment between our Concept-Graph model and a baseline ResNet-50. The left image shows the result of removing the most important regions—identified by our model—from the input of the baseline CNN. The right side shows the effect of adding those regions. The corresponding drop and rise in prediction confidence, respectively, suggest that our model localizes semantically meaningful regions similarly to the baseline CNN.

## 5.4 Qualitative Evaluation

In this section, we present visual examples to qualitatively demonstrate the interpretability of our proposed concept-based framework. Our goal is to assess whether the explanations are interpretable and whether they consistently behave as expected.

We first validate our approach visually using a subset of the ImageNet dataset. We then demonstrate the application of our visually grounded explanations to medical image datasets (*Derm7pt* and *PH<sup>2</sup>*). It is important to emphasize that our visual concepts are generated entirely without contextual supervision, such as clinical annotations in medical

datasets. Thus, our approach aims primarily at explaining model decisions from a purely visual perspective, rather than relying on clinically informed concepts.

### 5.4.1 General-Purpose Dataset

We begin by demonstrating concept generation on a general-purpose dataset using two semantically similar classes from ImageNet: *Ambulance* and *Recreational Vehicle*. Our goal here is to briefly provide a clear illustration of how our visual explanations work before moving to medical datasets, where visual concepts tend to be less easily discernible compared to the highly recognizable features of general objects.

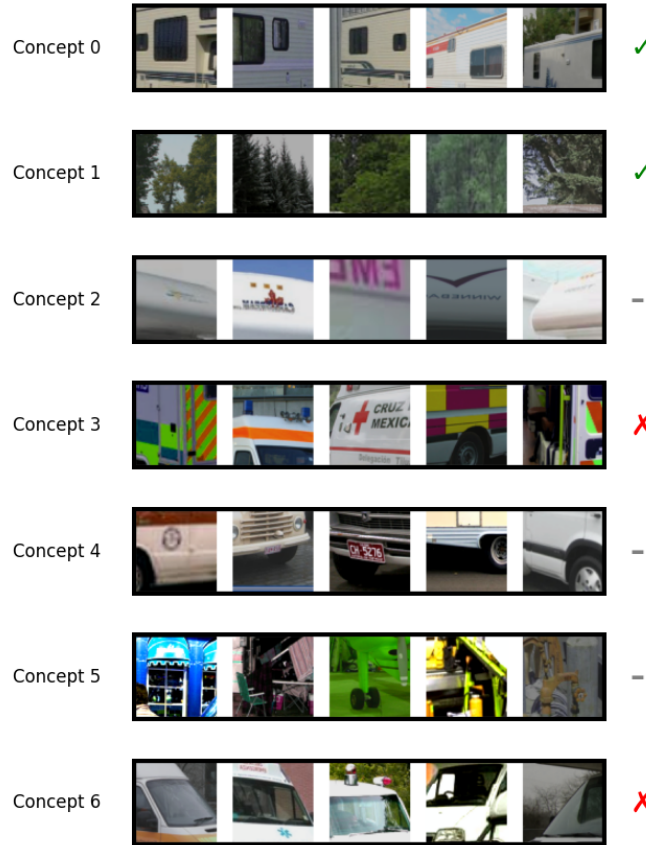


Figure 5.8: Top-5 representative samples per concept for the *Ambulance* and *Recreational Van* classes of ImageNet dataset. The dominant class for each concept is determined based on its *discriminativity score*. ✗ indicates concepts predominantly associated with *Ambulance*, ✓ indicates concepts predominantly associated with *Recreational Van*, and gray – indicates concepts evenly distributed across both classes. The average *discriminativity* value across discriminative concepts is approximately 0.80, meaning about 80% of activations for each discriminative concept belong to its associated class.

First we assess the generation of concept generation. Here the number of concepts was set to 7, based on the optimal *disentanglement score* defined in Section 4.4. As shown in Figure 5.8, the learned concepts exhibit a range of visually coherent and class-discriminative patterns. For instance, Concept 0 captures the characteristic side view of a

recreational vehicle, often showing doors and windows. Concept 1 highlights outdoor scenery such as greenery and trees, reflecting the typical context in which recreational vehicles are found. Both concepts are predominantly associated with the *Recreational Vehicle* class.

Concepts 2 and 4 contain more generic vehicle parts—such as hoods, bumpers, and tires—and are more ambiguous in nature, appearing across both classes. Similarly, Concept 5 depicts various equipment items (e.g., chairs, stretchers), which could plausibly be linked to either *Ambulance* or *Recreational Vehicle*, representing a semantically overlapping concept.

In contrast, the remaining concepts show strong alignment with the *Ambulance* class. Concept 3 captures symbolic patterns such as red crosses and emergency markings, which are visually characteristic of ambulances. Concept 6 focuses on the upper region of emergency vehicles, including the siren and light bar areas.

These visual cues are further supported by the *discriminativity* component of *disentanglement scores* of each concept, which indicate the class each concept predominantly originates from (see Figure 5.8). The analysis reveals that out of seven total concepts, four are class-discriminative while the remaining three are more generic. The average *discriminativity* value across the discriminative concepts is approximately 0.80, meaning that around 80% of activations for each of these concepts are associated with a single dominant class. These observations suggest that the generated concepts are generally coherent and class-discriminative. Even the ambiguous concepts appear to group visually similar patterns in a meaningful way, justifying the validity of visually grounded concept representations.

Having established that our learned concepts are visually coherent and class-discriminative, we now turn to a concrete example of how these concepts drive an individual prediction. Figure 5.9 illustrates the full explanation pipeline, including the model’s prediction, the most important image regions, the contribution of each concept, and representative examples for those concepts.

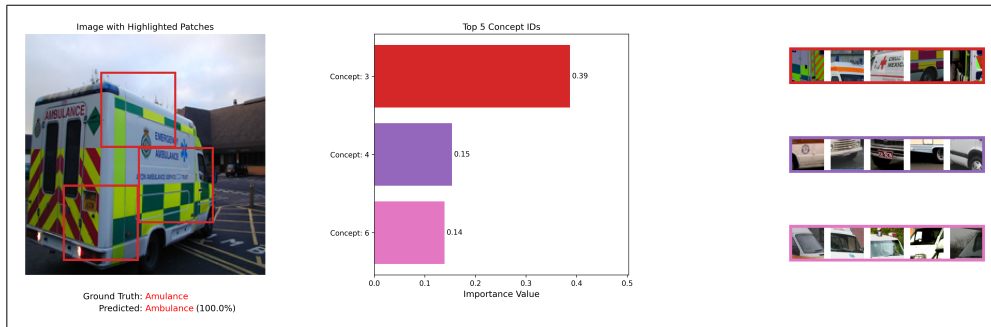


Figure 5.9: Concept-based explanation for an *Ambulance* image. **Left:** Input image with bounding boxes marking the three most important patches for the prediction; each bounding box is colour-coded according to the most active concept within that region. **Center:** Normalized contribution scores of the top three concepts; Concepts 3, 4, and 6 contribute 0.39, 0.15, and 0.14, respectively. **Right:** Representative samples for Concepts 3 (red), 4 (purple), and 6 (pink); the order and framing colours correspond exactly to the bars in the centre plot for concept importance. Below the visualization, we report the ground-truth label and the model’s prediction with confidence.

From Figure 5.9, we observe that:

- The top patches are centred on the ambulance’s distinctive symbols.
- Concept 3 (red) focuses on the emergency-symbol region (siren and cross markings) and has the highest contribution (0.39), strongly signalling the *Ambulance* class.
- Concept 4 (purple) captures more generic vehicle parts (e.g., bumpers, wheels), contributing 0.15.
- Concept 6 (pink) highlights the upper light (siren)-bar area, contributing 0.14.

This example demonstrates that our visually grounded explanations not only identify the correct class but also provide an intuitive decomposition of the model’s decision in terms of meaningful image regions and concepts. Overall, these qualitative results show that our model not only achieves accurate predictions but does so while remaining inherently interpretable, decomposing each decision into intuitive, visually grounded concepts.

### 5.4.2 Skin Lesion Dataset

Having validated that our concept generation and explanation pipeline produces interpretable and class-discriminative concepts on a general-purpose dataset, we now turn to the primary focus of our work: medical image datasets. Specifically, we evaluate our framework on *PH<sup>2</sup>* and *Derm7pt*, using the binary classification setting of *Melanoma* vs. *Nevus*.

Our central goal is to provide model interpretability in the medical domain using only visual modality, without relying on clinical annotations or expert-defined concepts. While quantitative results already demonstrated that our framework can train inherently interpretable models with good fidelity, here we qualitatively assess how well the learned concepts align with visually meaningful patterns in medical images.

#### Concept Generation Evaluation

Interpreting concepts in medical imaging is inherently more challenging than in general-purpose datasets, due to the subtler and more ambiguous nature of visual semantics. Despite this, our framework aims to extract interpretable, class-discriminative concepts that reflect clinically relevant visual cues. Based on the optimal *disentanglement score*, we selected 10 and 11 concepts for *PH<sup>2</sup>* and *Derm7pt*, respectively. This selection helps find a balance between semantic coherence and explanatory coverage, and also facilitates identifying which concepts are strongly class-aligned versus generic or ambiguous. Here, the higher optimal number of concepts selected for medical datasets (e.g., 10 for *PH<sup>2</sup>* and 11 for *Derm7pt*) reflect the need to capture a broader range of fine-grained semantic features in order to explain subtle distinctions in visual appearance. This aligns with the clinical setting that skin lesions often exhibit nuanced, overlapping patterns that demand more granular representations for meaningful interpretability.

Figure 5.10 shows the top-5 representative examples for each concept learned from the *PH<sup>2</sup>* dataset. As indicated in Table 5.1, 8 out of the 10 concepts are class-discriminative, with an average *discriminativity* value of approximately 75%, suggesting that a majority of the concept activations are associated with a single dominant class.

A closer inspection reveals that the learned concepts align well with visually meaningful cues commonly observed in clinical dermatology. Specifically, the concepts displayed in

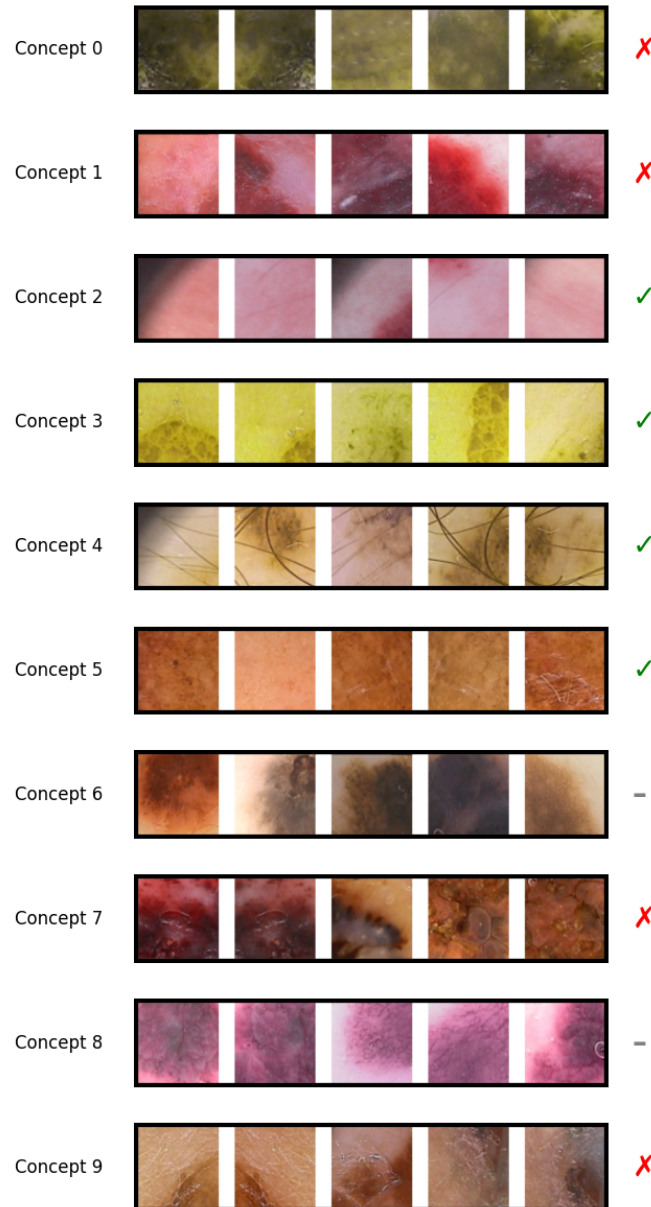


Figure 5.10: Top-5 representative samples per concept for the *Melanoma* and *Nevus* classes of the PH2 dataset. The dominant class for each concept is determined based on its *disentanglement score*. ✗ indicates concepts predominantly associated with *Melanoma*, ✓ indicates concepts predominantly associated with *Nevus*, and gray — indicates concepts evenly distributed across both classes. The average *discriminativity* value across discriminative concepts is approximately 0.75, meaning about 75% of activations for each discriminative concept belong to its associated class.

Figure 5.10 can be grouped based on their visual features and dominant class associations as follows:

### 1. Melanoma-associated concepts:

- **Concept 0:** Exhibits dark, irregular yellowish-brown pigmentation and a granular texture, indicative of atypical pigmentation and possible structural irregularities common in melanomas.
- **Concept 1:** Highlights bright-red, inflamed, or haemorrhagic regions, consistent with bleeding or ulceration often found in melanomas.
- **Concept 7:** Captures dark crust-like textures or scabby areas, possibly reflecting necrotic or ulcerated surfaces characteristic of advanced melanoma lesions.
- **Concept 9:** Represents dark brown or black pigmentation with irregular shapes and sharply defined edges, suggestive of malignant melanoma.

### 2. Nevus-associated concepts:

- **Concept 2:** Shows uniformly smooth, pinkish areas, indicative of healthy, benign skin texture and color distribution.
- **Concept 3:** Highlights yellowish honeycomb or reticular patterns, potentially associated with benign keratin or sebaceous features.
- **Concept 4:** Clearly captures fine hair growth through pigmented areas, a feature frequently seen in benign nevi, as melanoma rarely permits hair growth.
- **Concept 5:** Depicts consistent, evenly pigmented brownish areas with smooth texture and uniform appearance, typical of benign melanocytic lesions.

### 3. Ambiguous or generic concepts:

- **Concept 6:** Illustrates shadowed or darkened areas, common in both melanoma (due to depth and irregular pigmentation) and nevus (due to benign pigment clusters or curvature).
- **Concept 8:** Captures purplish-pink vascular or erythematous patterns that can appear in either benign or malignant lesions, likely related to lesion inflammation or vascularity.

These observations illustrate that, despite being unsupervised, our framework successfully uncovers coherent, visually interpretable patterns that align strongly with medically relevant features.

## Patch Localization and Concept Contribution

To further evaluate the interpretability of our framework, we examine the spatial localization of the most influential image regions contributing to the model's predictions. Unlike segmentation-based approaches, our method does not rely on pixel-level supervision. Instead, it infers importance based on the learned concept representations and their activation strengths.

Figure 5.11 presents qualitative examples, where the top three most relevant patches are highlighted per image using bounding boxes. These patches are selected based on their contribution to the final prediction via concept activation. Moreover, in our explanation pipeline each bounding box is colour-coded according to the most active concept within that region—allowing us not only to localize influential image areas but also to conceptually interpret what kind of visual pattern was detected there (see



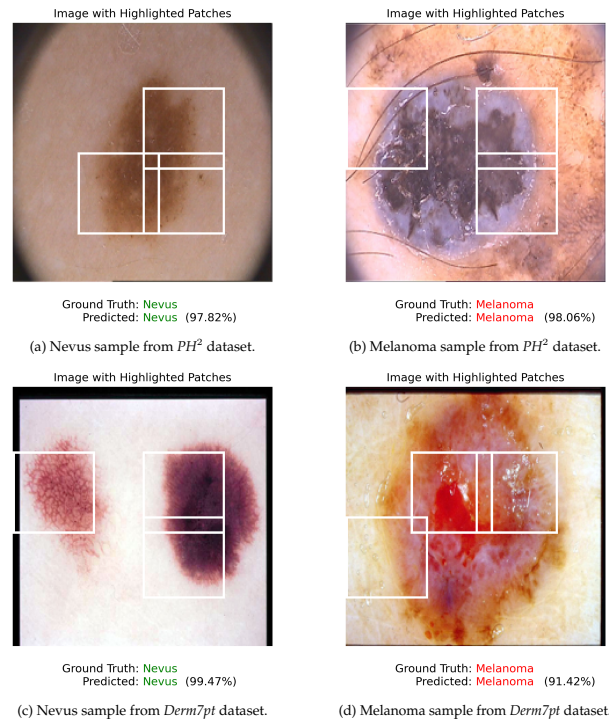


Figure 5.11: Patch-level localization for correctly classified samples from the  $PH^2$  and Derm7pt datasets. Each image is overlaid with bounding boxes highlighting the top three most influential patches identified by the model. These regions correspond to semantically meaningful areas of the lesion, demonstrating that the model focuses on clinically relevant visual cues without spatial supervision.

Figure 5.12). This makes the patch localization inherently semantic: each region can be associated with a specific concept (e.g., fine hair, keratin structures, or uniformly pigmented skin), further enhancing the explainability of the prediction.

Figure 5.12 moreover illustrates how the model leverages concept-based reasoning to correctly classify a *Nevus* sample. The highlighted patches are dominated by Concepts 4, 8, and 5, with contributions of 0.33, 0.12, and 0.10 respectively. These concepts correspond to:

- Concept 4 (purple): Associated with *Nevus*, capturing fine hair and benign keratin structures. It has the highest contribution, suggesting the model strongly relies on benign skin texture for this prediction.
- Concept 5 (brown): Also aligned with *Nevus*, highlighting smooth, uniformly pigmented skin. Its presence further reinforces a benign interpretation.
- Concept 8 (yellow): A more ambiguous concept appearing across both classes, showing purple-pinkish texture. While it contributes modestly (0.12), its neutral nature does not conflict with the overall *Nevus* prediction.

This result demonstrates that our model grounds its prediction in medically relevant visual patterns, such as uniform pigmentation and benign surface features. The high

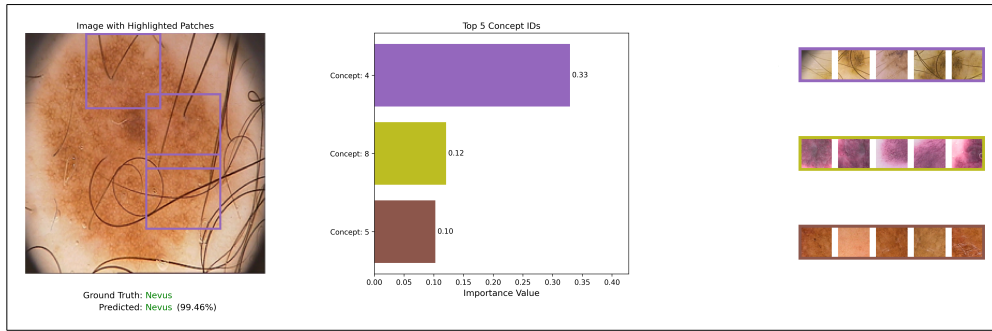


Figure 5.12: Concept-based explanation for a correctly classified *Nevus* image from the *PH²* dataset. **Left:** Input image with bounding boxes indicating the three most influential patches. **Center:** Contribution scores for the top three concepts. **Right:** Representative examples for Concepts 4 (purple), 8 (yellow), and 5 (brown). Bounding box colours correspond to the dominant concept activating in each patch.

influence of class-aligned concepts and the spatial focus on semantically meaningful patches highlight the interpretability and trustworthiness of the explanation pipeline.

### Case Study with Full Explanations

In this section, we observe some of the correct and wrong predictions assess how our model behaves in both scenarios. In addition to concept importance and spatial patch localization, here we further mark the concept-level examples with class-discriminative indicators. Specifically, each row of representative samples is prefixed with a ✓, ✗, or – symbol, denoting whether the corresponding concept is predominantly associated with *Nevus*, *Melanoma*, or is semantically ambiguous.

These additional markers provide helpful cues for interpretation, particularly in medical datasets where visual differences between classes are often subtle and difficult to discern. It is worth noting, however, that a concept being predominantly associated with one class does not imply complete disentanglement. Due to the additive nature of NMF, some degree of overlap or semantic mixing may still occur, especially in features shared across lesion types.

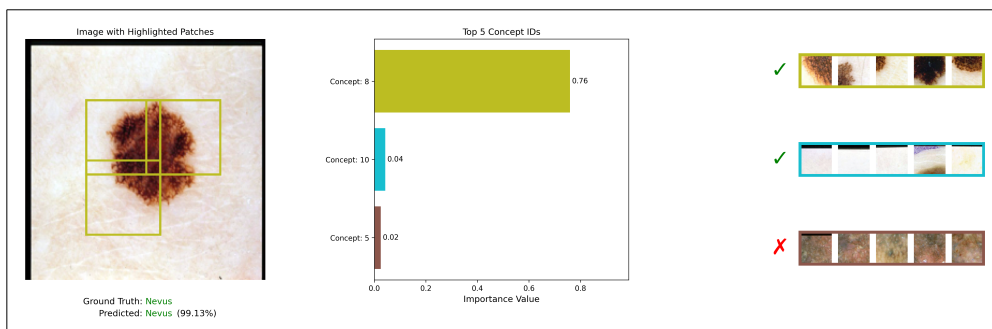


Figure 5.13: Concept-based explanation for a correctly classified *Nevus* image from the *Derm7pt* dataset. Bounding boxes indicate top contributing patches, colored by the most influential concept in each region.

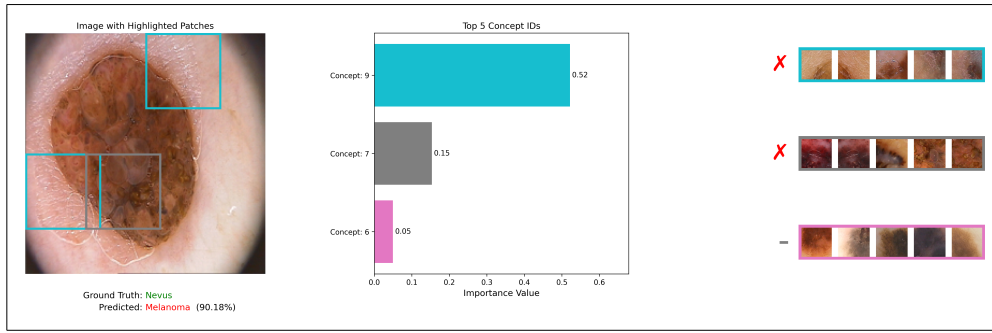


Figure 5.14: Concept-based explanation for a misclassified *Nevus* image from the  $PH^2$  dataset, predicted as *Melanoma*.

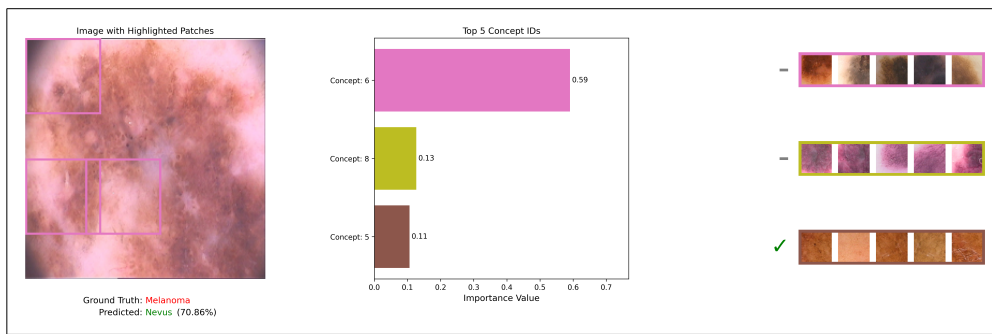


Figure 5.15: Concept-based explanation for a misclassified *Melanoma* image from the  $PH^2$  dataset, predicted as *Nevus*.

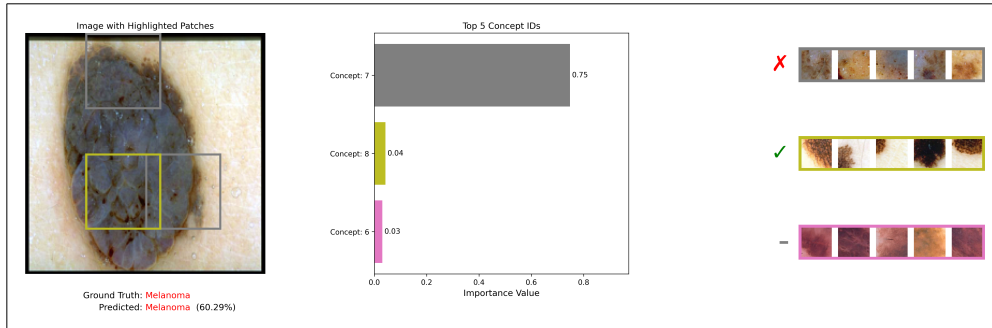


Figure 5.16: Correctly predicted *Melanoma* case from *Derm7pt* dataset using high-scoring, mixed-relevance concepts.

**Case 1 (Figure 5.13):** The model correctly classifies the *Nevus* image with high confidence. The top contributing concept is Concept 8 (✓) by 0.76, which captures dark, sharp-edged borders that often occur in benign uniform patterns. This concept is clearly dominant both in terms of contribution value and patch activation. Concepts 10 (✓) and 5 (✗) are minimally activated by less than 0.05. While Concept 5 is misaligned, its low influence renders it harmless in term of negative influence to the prediction. Overall, this is a consistent example of faithful prediction supported by class-aligned semantics.

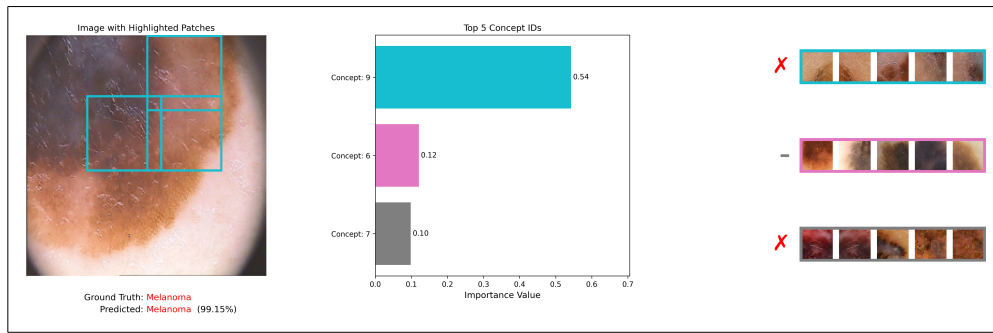


Figure 5.17: Correctly classified *Melanoma* image from  $PH^2$ , supported by well-aligned class-specific concepts.

**Case 2 (Figure 5.14):** This *Nevus* sample is misclassified as *Melanoma*, with strong activation of Concept 9 (X) and Concept 7 (X). Both are strongly associated with *Melanoma*, which explains the model’s mistake. The top patches correspond to irregular texture and ulcerated regions, which indeed resemble malignancy. While the prediction is incorrect, the model’s reasoning is not arbitrary — it reveals a known challenge in dermoscopy where some benign lesions present malignancy-like features, such as dark brown and crust-like textures.

**Case 3 (Figure 5.15):** In this example, the model fails to recognize a true *Melanoma*, predicting it instead as *Nevus*. Most important concepts, namely Concept 6 and Concept 8 are ambiguous (–), and Concept 5 (✓) contributes the most aligned benign signal. The highlighted patches are distributed over smooth and low-contrast areas, missing critical malignancy indicators. This suggests that the model may struggle when malignant signs are subtle.

**Case 4 (Figure 5.16):** The model correctly classifies this image as *Melanoma*, although the explanation reveals weakly mixed semantic clues. The dominant contributor is Concept 7 (X) by 0.75, which is appropriately associated with malignant structures and supports the prediction. Concept 8 (✓) is typically observed in benign cases but shares some visual similarity to edge features found in malignancies. However, its contribution is minimal (0.04), suggesting it does not significantly influence the model’s decision. This case illustrates how the model can produce a correct prediction despite partially relying on weakly relevant or semantically ambiguous features — a potential instance of compensatory reasoning across concepts.

**Case 5 (Figure 5.17):** This is a clear and well-aligned *Melanoma* prediction. The top activated concept is Concept 7 (✓), with additional support from Concept 8 (✓) and 6 (–). The highlighted patches correspond well to pigment irregularities and asymmetric structures typical in malignancy, indicating that the model attends to clinically relevant areas and leverages strong concept cues for explanation.

These qualitative examples illustrate both the strengths and limitations of our concept-based interpretability framework. When semantic signals are strong and well-aligned with class-specific concepts, the model produces confident and correctly justified predictions. In such cases, explanations are intuitive and the most influential patches clearly

correspond to clinically meaningful features. However, when features are subtle, shared across classes, or semantically ambiguous, the model can misattribute relevance—either making incorrect predictions or relying on weak or misleading cues. These cases underscore the importance of robust and disentangled concept learning, particularly in domains like medical imaging where visual differences are often nuanced.

## 5.5 Limitations

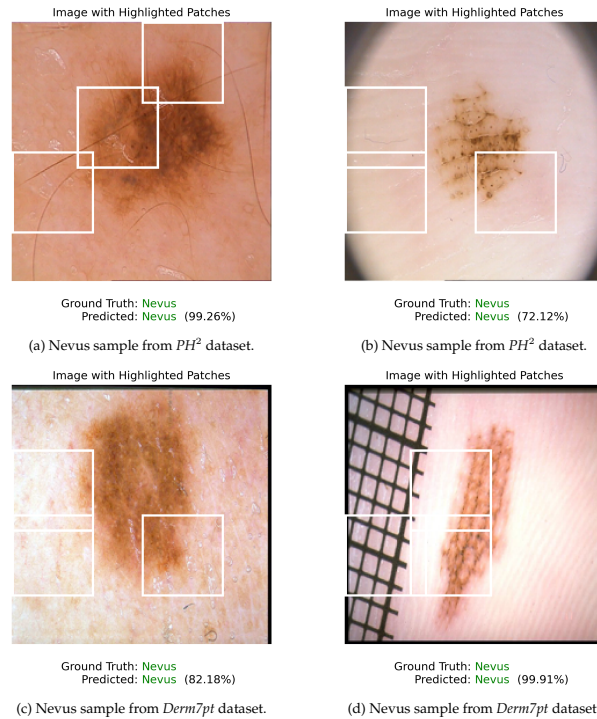


Figure 5.18: Patch-level localization examples where the model focuses on surrounding skin regions or unrelated visual artifacts instead of diagnostically relevant lesion areas. These cases highlight occasional misattribution of importance, potentially due to low-level visual biases.

We begin this section by examining common failure cases to better understand the limitations of the explanations generated by our model. In several instances, we observe that the model, despite arriving at the correct prediction, relies on irrelevant or weakly informative concepts. The top contributing concepts in these cases are often misaligned with the target lesion class, revealing possible vulnerabilities in the attribution process—particularly in scenarios involving subtle or fine-grained semantic features (see Figures 5.19 and 5.20). Additionally, as illustrated in Figure 5.18, the model sometimes attends to surrounding skin regions or unrelated artifacts, rather than the diagnostically relevant parts of the lesion, further highlighting the challenges of visual explanation in medical imaging contexts. However, this challenge can also be attributed to the difficulty of distinguishing certain visual cues in medical imaging, as many structures in the surrounding region appear visually similar. Consequently, the neighborhood context may substantially influence the classification decision [15].

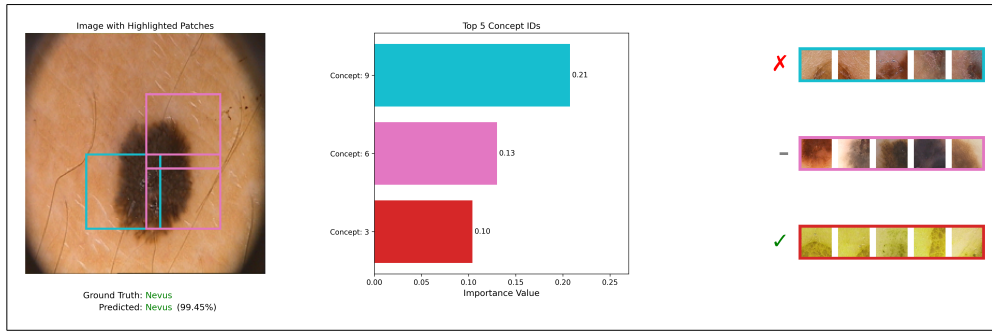


Figure 5.19: Failure case where the model makes a correct benign classification but the dominant concept is not strongly aligned with the target class. Although the highlighted patches focus on the lesion area, the top contributing concept (Concept 9, ✗) is generally associated with malignant artifacts, sharing only superficial visual similarities with the current lesion. This indicates the model’s difficulty in detecting fine-grained, semantically accurate features, even when the overall prediction is correct.

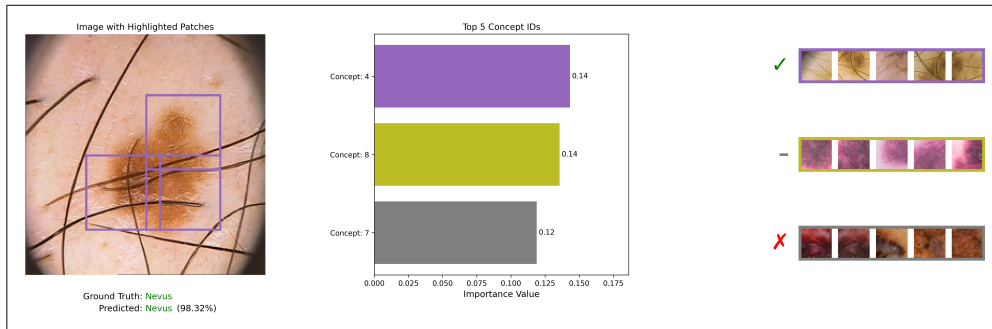


Figure 5.20: Failure case where the model correctly predicts benign class, but the explanation reveals a certain degree of semantic inconsistency. While the top-ranked concept (Concept 4, ✓) aligns well with the target class, a significant contribution also comes from Concept 7 (✗), which is typically associated with malignant features. This highlights limitations in concept disentanglement and suggests that the model may be sensitive to subtle, mixed visual cues.

These examples underscore the inherent challenges in generating reliable visual explanations for medical images, particularly when dealing with fine-grained or overlapping features. While our framework demonstrates promising classification performance on skin lesion datasets, several limitations should be acknowledged:

- Sensitivity to Subtle Features:** While the model can effectively leverage strong, class-specific visual signals, it may fail to detect more subtle yet important cues for its explanations. In some cases, these features are misinterpreted or confused with visually similar patterns associated with the opposing class, potentially degrading both predictive reliability and explanation quality. *We observed cases where concepts with only superficial similarity contributed to correct predictions, raising concerns about robustness under fine-grained distinctions.*
- Concept Drift Across Samples:** Some concepts, particularly the ambiguous ones, appear across both benign and malignant classes. Although they often have low



contribution scores, their presence may indicate latent concept drift or class overlap, which can impact both prediction stability and trustworthiness. *In some cases, even class-opposing concepts were attributed significant influence, reflecting limitations in semantic disentanglement.*

- **Lack of Pixel-Level Supervision:** Our method operates without pixel-wise or region-level annotations, relying solely on weak supervision via image-level labels. While this enables unsupervised concept discovery, it may result in coarse or noisy localization, especially when lesions span large or heterogeneous regions. *We observed instances where the model focused on background or non-lesion regions, suggesting noise in patch importance attribution.*
- **Loss of Semantic Coherence in Patch Pooling:** Our patch-based concept modelling relies on spatial decomposition of the input image into local regions. While this enables patch-based analysis, it can also disrupt the semantic continuity of larger structures, especially when meaningful patterns span across patch boundaries. As a result, semantically coherent features may be fragmented during pooling, leading to diluted or ambiguous concept representations.
- **Suboptimal Representation:** Small sample size, along with class imbalance, can lead to weaker and less diverse concept generation. In other words, concept basis generation methods (e.g., NMF) may fail to learn rich representations due to limited information. Although data augmentation is often used to address such issues, small visual perturbations may introduce noise that disrupts the matrix decomposition process, ultimately degrading concept coherence. Therefore, the perturbation choices during concept generation must be curated carefully. *Overall, a small sample size may lead to explanations that rely on less informative or even misleading visual cues.*

Despite the promising predictive performance and interpretability offered by our concept-based approach, several limitations remain. Small sample sizes and class imbalance hinder the diversity and granularity of the learned concept representations. Furthermore, due to the additive nature of NMF, a concept being predominantly associated with a class does not imply full disentanglement; overlapping or mixed semantics may still emerge, especially for features shared across lesion types. Additionally, the spatial patch breakdown used for localization can fragment semantically meaningful structures, potentially reducing coherence in the extracted concepts.

Overall explanation produced by our pipeline exhibits a certain degree of instability particularly when subtle semantic cues are involved. Our results show that the proposed approach holds potential for enhancing transparency in medical imaging; however, its current limitations must be addressed, and the generated explanations should be rigorously validated before deployment in real-world clinical contexts.

---

## Chapter 6

# Conclusion

In this thesis, we began by providing a historical perspective on artificial intelligence and explainable artificial intelligence [19], underscoring the critical importance of interpretability, especially in high-stakes fields such as healthcare. We outlined the necessity of clear, trustworthy explanations, highlighting the limitations inherent in current attribution-based explanation methods [1], and thereby motivating a shift toward concept-based explainability. To provide a comprehensive foundation, we conducted a detailed review of existing concept-based methods, illustrating their strengths and shortcomings, particularly the dependency on textual or clinical annotations [57, 83].

Building upon this foundation, we introduced an inherently interpretable, visually grounded concept-based explainability framework specifically designed for medical image analysis. Our proposed approach bridges the gap between high-performing deep learning models and clinical interpretability without relying on textual or clinical annotations. Leveraging Non-negative Matrix Factorization, we developed an unsupervised concept discovery method that extracts visually coherent and semantically meaningful concepts directly from images. These concepts serve as foundational elements for structured, patch-based concept graphs, processed through Graph Attention Networks. By combining patch-level pooling strategies and attention mechanisms, our framework facilitates fine-grained localization and relational reasoning of visual features, significantly enhancing the model’s interpretability and transparency.

We thoroughly evaluated our framework across diverse datasets, including both general-purpose (*ImageNet*) and medical imaging datasets (*HAM10000*, *Derm7pt*, *PH<sup>2</sup>*). Our results demonstrated that our visually grounded approach achieves competitive classification performance relative to supervised concept-based methods, while providing interpretable, concept-level explanations with high fidelity and precise localization.

Despite promising predictive performance, we identified several limitations, such as potential loss of contextual information due to patch pooling strategies, concept drift across samples, and challenges in effectively handling subtle visual features and overlapping semantics. To address these issues, we propose the following promising directions for future research:



## 6.1 Future Work

Building on the foundation of our current framework, future research may explore several key directions to enhance its effectiveness and applicability:

- **Context-Aware Patch Pooling:** When an image is divided into patches, critical spatial context and global semantics may be lost, which can hinder interpretability. Future work could explore hierarchical or spatially-aware pooling strategies that preserve relationships between patches. In particular, spatial attention mechanisms [116, 62] that consider neighbouring regions and their interactions may help maintain contextual integrity and enhance explanation quality.
- **Improved Concept Disentanglement:** Exploring better concept generation techniques and integrating supervised or semi-supervised learning approaches could yield more robust and semantically distinct concept representations, thereby reducing concept drift and semantic overlap.
- **Human-in-the-Loop Frameworks:** Incorporating expert knowledge via interactive, human-in-the-loop interfaces would significantly enhance the validation and refinement of discovered concepts [53, 6]. Such frameworks enable domain experts to iteratively guide and improve the concept discovery process, ensuring greater clinical relevance and interpretability.
- **Interactive Explanation Tools:** Developing user-friendly, interactive explanation frameworks represents a valuable area of improvement. Although existing interactive tools provide limited functionalities, enriched interfaces could facilitate active exploration, improve contextual awareness, and deliver personalized insights [50]. By leveraging the additive, permutation-invariant nature of our concept representation, interactive exploration can be performed without introducing perturbation bias, enabling dynamic, real-time manipulation of patch and concept-level information. Richer interaction capabilities would further enhance interpretability and transparency by supporting active exploration, contextual understanding, and personalized explanations.

In conclusion, our research represents a promising step toward more transparent, trustworthy, and inherently interpretable models for medical imaging. By addressing the identified limitations and embracing the suggested improvements, we aim to enhance the clinical adoption and reliability of concept-based explainability methods, ultimately fostering greater trust and efficacy in AI-driven healthcare solutions.

---

# Bibliography

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2020. Sanity Checks for Saliency Maps. (2020). <https://arxiv.org/abs/1810.03292>
- [2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805. DOI:<http://dx.doi.org>/<https://doi.org/10.1016/j.inffus.2023.101805>
- [3] Sule T. Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Adaptive Agents and Multi-Agent Systems*. <https://api.semanticscholar.org/CorpusID:169034726>
- [4] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker, and Catherine Mooney. 2021. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences* 11 (2021), 5088. <https://api.semanticscholar.org/CorpusID:236420019>
- [5] Yoshua Bengio. 2012. Practical recommendations for gradient-based training of deep architectures. (2012). <https://arxiv.org/abs/1206.5533>
- [6] Matteo Bianchi, Antonio De Santis, Andrea Tocchetti, and Marco Brambilla. 2024. Interpretable Network Visualizations: A Human-in-the-Loop Approach for Post-hoc Explainability of CNN-based Image Classification. (2024). <https://arxiv.org/abs/2405.03301>
- [7] Yequan Bie, Luyang Luo, and Hao Chen. 2024. MICA: Towards Explainable Skin Lesion Diagnosis via Multi-Level Image-Concept Alignment. (2024). <https://arxiv.org/abs/2401.08527>
- [8] Christopher M. Bishop and Hugh Bishop. 2024. *Deep Learning: Foundations and Concepts*. Springer. DOI:<http://dx.doi.org/10.1007/978-3-031-45468-4>
- [9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. (2020). <https://arxiv.org/abs/2005.14165>

- [10] Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-Yan Liu, and Liwei Wang. 2021. GraphNorm: A Principled Approach to Accelerating Graph Neural Network Training. (2021). <https://arxiv.org/abs/2009.03294>
- [11] Diogo Vieira Carvalho, Eduardo Marques Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* (2019). <https://api.semanticscholar.org/CorpusID:199659548>
- [12] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. DOI:<http://dx.doi.org/10.1109/wacv.2018.00097>
- [13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. (2017). <https://arxiv.org/abs/1706.05587>
- [14] Zhi Chen, Yijie Bei, and Cynthia Rudin. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence* 2, 12 (Dec. 2020), 772–782. DOI: <http://dx.doi.org/10.1038/s42256-020-00265-z>
- [15] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. 2013. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 411–418.
- [16] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). (2016). <https://arxiv.org/abs/1511.07289>
- [17] Julien Colin, Thomas Fel, Remi Cadene, and Thomas Serre. 2023. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. (2023). <https://arxiv.org/abs/2112.04417>
- [18] European Commission, Content Directorate-General for Communications Networks, Technology, and Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji. 2019. *Ethics guidelines for trustworthy AI*. Publications Office. DOI: <http://dx.doi.org/doi/10.2759/346720>
- [19] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning* (2017). <https://api.semanticscholar.org/CorpusID:11319376>
- [20] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?. In *Findings of the Association for Computational Linguistics: EACL 2023*. 1151–1163.
- [21] Thomas Fel. 2025. Sparks of Explainability: Recent Advancements in Explaining Large Vision Models. (2025). <https://arxiv.org/abs/2502.01048>
- [22] Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Léo andéol, Mathieu Chalvidal, and Thomas Serre. 2023. A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation. (2023). <https://arxiv.org/abs/2306.07304>

- [23] Thomas Fel, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. 2021. Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. (2021). <https://arxiv.org/abs/2111.04138>
- [24] Thomas Fel, Melanie Ducoffe, David Vigouroux, Remi Cadene, Mikael Capelle, Claire Nicodeme, and Thomas Serre. 2023a. Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis. (2023). <https://arxiv.org/abs/2202.07728>
- [25] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. 2023b. CRAFT: Concept Recursive Activation FacTorization for Explainability. (2023). <https://arxiv.org/abs/2211.10154>
- [26] H. L. G., F. Flammini, R. Kumar V, and P. N. S. (Eds.). 2025. *Recent Trends in Healthcare Innovation: Proceedings of the Annual International Conference on Recent Trends in Healthcare Innovation (AICRTHI 2024), Mysuru, India, October 24th–25th, 2024* (1 ed.). CRC Press. DOI:<http://dx.doi.org/10.1201/9781003501367>
- [27] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li. 2023. A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. (2023). <https://arxiv.org/abs/2109.12843>
- [28] Yunhao Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyang Wu. 2021. A Peek Into the Reasoning of Neural Networks: Interpreting with Structural Visual Concepts. (2021). <https://arxiv.org/abs/2105.00290>
- [29] Sahra Ghalebikesabi, Lucile Ter-Minassian, Karla Diaz-Ordaz, and Chris Holmes. 2021. On Locality of Local Explanation Models. (2021). <https://arxiv.org/abs/2106.14648>
- [30] Amirata Ghorbani, Abubakar Abid, and James Zou. 2018. Interpretation of Neural Networks is Fragile. (2018). <https://arxiv.org/abs/1710.10547>
- [31] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. (2019). <https://arxiv.org/abs/1902.03129>
- [32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. Available online at <http://www.deeplearningbook.org>.
- [33] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision Making and a “Right to Explanation”. *AI Magazine* 38, 3 (Sept. 2017), 50–57. DOI:<http://dx.doi.org/10.1609/aimag.v38i3.2741>
- [34] Mara Graziani, Vincent Andrearczyk, and Henning Müller. 2019. Regression Concept Vectors for Bidirectional Explanations in Histopathology. (2019). <https://arxiv.org/abs/1904.04520>
- [35] Mara Graziani, An phi Nguyen, Laura O’Mahony, Henning Müller, and Vincent Andrearczyk. 2023. Concept discovery and Dataset exploration with Singular Value Decomposition. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*. <https://openreview.net/forum?id=iO1YmD1PtC8>

- [36] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. (2021). <https://arxiv.org/abs/2104.09957>
- [37] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)* 51 (2018), 1 – 42. <https://api.semanticscholar.org/CorpusID:3342225>
- [38] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? (2020). <https://arxiv.org/abs/2005.01831>
- [39] Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. 2021. On Baselines for Local Feature Attributions. (2021). <https://arxiv.org/abs/2101.00905>
- [40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (2015). <https://arxiv.org/abs/1512.03385>
- [41] Dan Hendrycks and Kevin Gimpel. 2023. Gaussian Error Linear Units (GELUs). (2023). <https://arxiv.org/abs/1606.08415>
- [42] Galib Muhammad Shahriar Himel, Md. Masudul Islam, Kh Abdullah Al-Aff, Shams Ibne Karim, and Md. Kabir Uddin Sikder. 2024. Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-based Non-invasive Digital System. (2024). <https://arxiv.org/abs/2401.04746>
- [43] Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl. 2023. *Transparency in Artificial Intelligence: Challenges and Perspectives*. Springer Nature Switzerland, Cham. DOI:<http://dx.doi.org/10.1007/978-3-031-63797-1>
- [44] Junlin Hou, Jilan Xu, and Hao Chen. 2024. Concept-Attention Whitening for Interpretable Skin Lesion Diagnosis. (2024). <https://arxiv.org/abs/2404.05997>
- [45] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. 2021. Evaluations and Methods for Explanation through Robustness Analysis. (2021). <https://arxiv.org/abs/2006.00442>
- [46] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. (2018). <https://arxiv.org/abs/1608.06993>
- [47] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015). <https://arxiv.org/abs/1502.03167>
- [48] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2020. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2020). <https://api.semanticscholar.org/CorpusID:222379602>

- [49] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2022. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (June 2022), 11945–11953. DOI:<http://dx.doi.org/10.1609/aaai.v36i11.21452>
- [50] Md Abdul Kadir, Abdulrahman Mohamed Selim, Michael Barz, and Daniel Sonntag. 2023a. A User Interface for Explaining Machine Learning Model Explanations. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 59–63. DOI:<http://dx.doi.org/10.1145/3581754.3584131>
- [51] Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. 2023b. Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*. 000111–000124. DOI:<http://dx.doi.org/10.1109/INES59282.2023.10297629>
- [52] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 2668–2677. <https://proceedings.mlr.press/v80/kim18d.html>
- [53] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the Human Interpretability of Visual Explanations. (2022). <https://arxiv.org/abs/2112.03184>
- [54] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. (2017). <https://arxiv.org/abs/1412.6980>
- [55] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. (2017). <https://arxiv.org/abs/1609.02907>
- [56] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:13193974>
- [57] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. (2020). <https://arxiv.org/abs/2007.04612>
- [58] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 1097–1105.
- [59] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.
- [60] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. DOI:<http://dx.doi.org/10.1109/5.726791>

- [61] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791. DOI:<http://dx.doi.org/10.1038/44565>
- [62] Yawei Li, Kai Zhang, Jiezhong Cao, Radu Timofte, Michele Magno, Luca Benini, and Luc Van Gool. 2025. LocalViT: Analyzing Locality in Vision Transformers. (2025). <https://arxiv.org/abs/2104.05707>
- [63] Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network In Network. (2014). <https://arxiv.org/abs/1312.4400>
- [64] Zachary C. Lipton. 2017a. The Doctor Just Won’t Accept That! (2017). <https://arxiv.org/abs/1711.08037>
- [65] Zachary C. Lipton. 2017b. The Mythos of Model Interpretability. (2017). <https://arxiv.org/abs/1606.03490>
- [66] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. 2020. On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–10. DOI: <http://dx.doi.org/10.1109/ijcnn48605.2020.9206946>
- [67] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. 2022. ExAID: A Multimodal Explanation Framework for Computer-Aided Diagnosis of Skin Lesions. (2022). <https://arxiv.org/abs/2201.01249>
- [68] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013), Workshop on Deep Learning for Audio, Speech and Language Processing*. [https://ai.stanford.edu/~amaas/papers/relu\\_hybrid\\_icml2013\\_final.pdf](https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf)
- [69] Anna Majkowska, Sid Mittal, David F. Steiner, Joshua Jay Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, Alexander Ding, Greg S Corrado, Daniel Tse, and Shravya Shetty. 2019. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology* (2019), 191293. <https://api.semanticscholar.org/CorpusID:208611383>
- [70] Ričards Marcinkevičs, Patricia Reis Wolfertstetter, Ugne Klimiene, Kieran Chin-Cheong, Alyssia Paschke, Julia Zerres, Markus Denzinger, David Niederberger, Sven Wellmann, Ece Ozkan, Christian Knorr, and Julia E. Vogt. 2024. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Medical Image Analysis* 91 (Jan. 2024), 103042. DOI:<http://dx.doi.org/10.1016/j.media.2023.103042>
- [71] J. McCarthy, M. Minsky, N. Rochester, and C.E. Shannon. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine* 27 (12 2006).
- [72] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. *ArXiv* abs/1706.07269 (2017). <https://api.semanticscholar.org/CorpusID:36024272>

- [73] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. 2022. *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*. Springer International Publishing, Cham, 39–68. DOI: [http://dx.doi.org/10.1007/978-3-031-04083-2\\_4](http://dx.doi.org/10.1007/978-3-031-04083-2_4)
- [74] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2017. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* 73 (2017), 1–15. <https://api.semanticscholar.org/CorpusID:207170725>
- [75] Andreia Moraes, Jaime S. Marques, and Jorge Rozeira. 2013. PH<sup>2</sup> - A dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 5437–5440. DOI:<http://dx.doi.org/10.1109/EMBC.2013.6605014>
- [76] Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 4602–4609. DOI:<http://dx.doi.org/10.1609/aaai.v33i01.33014602>
- [77] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)*. Omnipress, Madison, WI, USA, 807–814.
- [78] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2019. Understanding Neural Networks via Feature Visualization: A survey. (2019). <https://arxiv.org/abs/1904.08939>
- [79] Duy M. H. Nguyen, Hoang Nguyen, Nghiem T. Diep, Tan N. Pham, Tri Cao, Binh T. Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, Daniel Sonntag, and Mathias Niepert. 2023. LVM-Med: Learning Large-Scale Self-Supervised Vision Models for Medical Imaging via Second-order Graph Matching. (2023). <https://arxiv.org/abs/2306.11925>
- [80] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2022. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. (2022). <https://arxiv.org/abs/2105.14944>
- [81] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. Label-Free Concept Bottleneck Models. (2023). <https://arxiv.org/abs/2304.06129>
- [82] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* (2017). DOI:<http://dx.doi.org/10.23915/distill.00007> <https://distill.pub/2017/feature-visualization>.
- [83] Cristiano Patrício, João C. Neves, and Luís F. Teixeira. 2023a. Coherent Concept-based Explanations in Medical Image and Its Application to Skin Lesion Diagnosis. (2023). <https://arxiv.org/abs/2304.04579>
- [84] Cristiano Patrício, João C. Neves, and Luís F. Teixeira. 2023b. Explainable Deep Learning Methods in Medical Image Classification: A Survey. (2023). <https://arxiv.org/abs/2205.04766>



- [85] Cristiano Patrício, Luís F. Teixeira, and João C. Neves. 2025. A two-step concept-based approach for enhanced interpretability and trust in skin lesion diagnosis. *Computational and Structural Biotechnology Journal* 28 (2025), 71–79. DOI:<http://dx.doi.org/10.1016/j.csbj.2025.02.013>
- [86] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. (2018). <https://arxiv.org/abs/1806.07421>
- [87] Naveen Raman, Mateo Espinosa Zarlenga, Juyeon Heo, and Mateja Jamnik. 2024. Do Concept Bottleneck Models Respect Localities? (2024). <https://arxiv.org/abs/2401.01259>
- [88] Bin Ren, Laurent Pueyo, Christine Chen, Élodie Choquet, John H. Debes, Gaspard Duchêne, François Ménard, and Marshall D. Perrin. 2023. Using Data Imputation for Signal Separation in High-contrast Imaging. (Aug. 2023). DOI:<http://dx.doi.org/10.3847/1538-4357/ab7024>
- [89] Veronica Rotemberg, Nathan Kurtansky, Brett Betz-Stablein, Liam Caffery, Elias Chousakos, Noel C. Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Andre Kalloo, Konstantinos Liopyris, Michael Marchetti, Ashfaq A. Marghoob, Scott Menzies, Nabin Mishra, Harald Kittler, Peter Soyer, and Philipp Tschandl. 2021. A dataset for the structured evaluation of dermoscopic image analysis algorithms. *Computerized Medical Imaging and Graphics* 88 (2021), 101820. DOI:<http://dx.doi.org/10.1016/j.compmedimag.2020.101820>
- [90] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. (2019). <https://arxiv.org/abs/1811.10154>
- [91] Waddah Saeed and Christian Walter Peter Omlin. 2021. Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. *ArXiv abs/2111.06420* (2021). <https://api.semanticscholar.org/CorpusID:244102736>
- [92] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. (2019). <https://arxiv.org/abs/1801.04381>
- [93] Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N Balasubramanian. 2021. A Framework for Learning Ante-hoc Explainable Models via Concepts. (2021). <https://arxiv.org/abs/2108.11761>
- [94] Simon Schrodi, Julian Schur, Max Argus, and Thomas Brox. 2024. Concept Bottleneck Models Without Predefined Concepts. (2024). <https://arxiv.org/abs/2407.03921>
- [95] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct. 2019), 336–359. DOI:<http://dx.doi.org/10.1007/s11263-019-01228-7>

- [96] Hua Shen and Ting-Hao Kenneth Huang. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. (2020). <https://arxiv.org/abs/2008.11721>
- [97] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2014). <https://arxiv.org/abs/1312.6034>
- [98] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. (2020). <https://arxiv.org/abs/1911.02508>
- [99] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. (2017). <https://arxiv.org/abs/1706.03825>
- [100] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. 2020. Visualizing the Impact of Feature Attribution Baselines. *Distill* (2020). DOI:<http://dx.doi.org/10.23915/distill.00022> <https://distill.pub/2020/attribution-baselines>.
- [101] Yiyu Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. (2022). <https://arxiv.org/abs/2204.06507>
- [102] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, 3319–3328. <https://proceedings.mlr.press/v70/sundararajan17a.html>
- [103] DGL Team. 2023. Deep Graph Library (DGL). <https://www.dgl.ai>. (2023). Version 2.0.x.
- [104] Ker Than. 2019. Ancient myths reveal early fantasies about creating artificial life. (2019). <https://news.stanford.edu/stories/2019/02/ancient-myths-reveal-early-fantasies-artificial-life> Accessed: 2025-05-19.
- [105] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude. [http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf). (2012). Coursera: Neural Networks for Machine Learning.
- [106] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5, 1 (2018), 180161. DOI:<http://dx.doi.org/10.1038/sdata.2018.161>
- [107] Michael van Lent, William Fisher, and Michael Mancuso. 2004. An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. In *AAAI Conference on Artificial Intelligence*. <https://api.semanticscholar.org/CorpusID:7286175>
- [108] Gal Vardi and Ohad Shamir. 2020. Neural Networks with Small Weights and Depth-Separation Barriers. (2020). <https://arxiv.org/abs/2006.00625>

- [109] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. (2023). <https://arxiv.org/abs/1706.03762>
- [110] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. (2018). <https://arxiv.org/abs/1710.10903>
- [111] Cédric Villani. 2009. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, Vol. 338. Springer.
- [112] Matthew J. Vowels. 2022. Trying to Outrun Causality with Machine Learning: Limitations of Model Explainability Techniques for Identifying Predictive Variables. (2022). <https://arxiv.org/abs/2202.09875>
- [113] Hongmei Wang, Junlin Hou, and Hao Chen. 2024. Concept Complement Bottleneck Model for Interpretable Medical Image Diagnosis. (2024). <https://arxiv.org/abs/2410.15446>
- [114] Zhu Wang, Yang Song, Shih-Cheng Huang Zhang, Hao Tang, Wenjia Huang, Lei Xing, Xiaowei Wang, Jing Xiao, and Le Lu. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *arXiv preprint arXiv:2211.14848* (2022).
- [115] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. 2022. Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement. *ArXiv abs/2203.08008* (2022). <https://api.semanticscholar.org/CorpusID:247450840>
- [116] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. (2018). <https://arxiv.org/abs/1807.06521>
- [117] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li, Naimeng Yao, Xinyi Wang, Xiaotong Gu, Muhammad Amin, and Byeong Kang. 2023a. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems* 3 (08 2023), 161 – 188. DOI:<http://dx.doi.org/10.1007/s44230-023-00038-y>
- [118] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li, Naimeng Yao, Xinyi Wang, Xiaotong Gu, Muhammad Bilal Amin, and Byeong Kang. 2023b. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems* 3, 3 (2023), 161–188. DOI: <http://dx.doi.org/10.1007/s44230-023-00038-y>
- [119] Mert Yuksekgonul, Maggie Wang, and James Zou. 2023. Post-hoc Concept Bottleneck Models. (2023). <https://arxiv.org/abs/2205.15480>
- [120] Zhiwei Zeng. 2022. Explainable artificial intelligence (XAI) for healthcare decision-making. (2022). DOI:<http://dx.doi.org/10.32657/10356/155849>
- [121] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. 2021. Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. (2021). <https://arxiv.org/abs/2006.15417>

- [122] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew Lungren, Tristan Naumann, and Hoifung Poon. 2023. Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing. (2023). DOI:<http://dx.doi.org/10.48550/ARXIV.2303.00915>

---

## Appendix A

### Disclosure

During the preparation of this thesis, I used generative digital tools to improve the clarity and academic quality of the written content. Specifically:

- **ChatGPT (OpenAI, GPT-4, 2023 release)** The tool was employed primarily for rephrasing my own text, improving grammar, and enhancing clarity and flow. In the sections Introduction, Related Work, and Technical Background, AI assistance was mainly limited to language polishing; in the sections Methodology, Experiments and Evaluation, and Conclusion, it was also used for summarising draft content I had written. The same tool was also used to clean and refactor my code to improve readability and reusability. The ideas and implementation of the code is my own. After the clean-up, I carefully verified that the code was properly refactored and that no unintended changes were introduced.
- **DeepL Write (DeepL, Beta, 2023 release)** was used to enhance grammar, sentence structure, and overall style beyond basic spelling or grammar correction.

Overall, these tools were used as writing/clean-up assistants and did not contribute to the generation of original research ideas, methodology, or results. I reviewed and approved every AI-assisted change.