Discovering Concept Directions from Diffusion-based Counterfactuals via Latent Clustering

Payal Varshney[a,b,*], Adriano Lucieri[a,b], Christoph Balada[a,b], Andreas Dengel[a,b], Sheraz Ahmed[b]

[a]*Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Gottlieb-Daimler-Street 47, Kaiserslautern, 67663, Rhineland-Palatinate, Germany*
[b]*German Research Center for Artificial Intelligence GmbH (DFKI), Trippstadter Str 122, Kaiserslautern, 67663, Rhineland-Palatinate, Germany*

Abstract

Concept-based explanations have emerged as an effective approach within Explainable Artificial Intelligence, enabling interpretable insights by aligning model decisions with human-understandable concepts. However, existing methods rely on computationally intensive procedures and struggle to efficiently capture complex, semantic concepts. This work introduces the Concept Directions via Latent Clustering (CDLC), which extracts global, class-specific concept directions by clustering latent difference vectors derived from factual and diffusion-generated counterfactual image pairs. CDLC reduces storage requirements by ∼4.6× and accelerates concept discovery by ∼5.3× compared to the baseline method, while requiring no GPU for clustering, thereby enabling efficient extraction of multidimensional semantic concepts across latent dimensions. This approach is validated on a real-world skin lesion dataset, demonstrating that the extracted concept directions align with clinically recognized dermoscopic features and, in some cases, reveal dataset-specific biases or unknown biomarkers. These results highlight that CDLC is interpretable, scalable, and applicable across high-stakes domains and diverse data modalities.

Keywords: Explainability, Counterfactual Explanation, Concept-Based Explanation, Latent Diffusion Model, Dermoscopy, Concept Directions, Clustering

1. Introduction

In high-stakes applications, such as medical diagnosis, financial risk assessment, and autonomous driving, understanding the rationale behind a neural network's decision is often as important as the decision itself. Explainable Artificial Intelligence (XAI) [1, 2] has emerged as a critical research area, aiming to bridge the gap between high-performing black-box models and human interpretability. Among the various XAI paradigms, concept-based explanations [3, 4] have gained particular attention due to their ability to express model behavior in terms of high-level, semantically meaningful concepts, rather than low-level feature weights or pixel-based saliency maps [5, 6]. By aligning explanations with concepts recognized by domain experts, these methods facilitate trust [7, 8], debugging [9], and regulatory compliance [10, 11].

Although concept-based explainability has been widely explored using convolutional [3, 4] and GAN-based architectures [12, 13], its application within diffusion-based generative models remains relatively underexplored. Recent works, such as Concept-Guided Latent Diffusion Counterfactual Explanations (CoLa-DCE) [14], have demonstrated the ability to produce spatially constrained, concept-conditioned counterfactuals for an arbitrary classifier. However, this approach only utilizes concepts to guide counterfactual explanations and does not extract global concept representations that are generalizable across examples.

Varshney et al. [15] proposed the Concept Discovery through Latent Diffusion-based Counterfactual Trajectories (CDCT) framework to discover global concepts that generalize across multiple samples. CDCT generates a classifier-guided counterfactual image trajectory dataset using a latent diffusion model [16] and subsequently trains a Variational Autoencoder (VAE) [17] on this trajectory dataset to disentangle classifier-relevant concepts. While CDCT represents a significant advancement in leveraging diffusion-based counterfactuals for concept discovery, it relies on a dimension-wise traversal strategy, wherein each latent variable is modified independently to detect relevant concepts. This exhaustive search procedure is computationally expensive, particularly in high-dimensional latent spaces, and inherently overlooks semantic concepts that arise from interactions among multiple latent dimensions. Consequently, CDCT often identifies only simple concepts, restricting its capacity to uncover high-level semantic directions.

To address these limitations, this paper proposes a novel framework Concept Directions via Latent Clustering (CDLC), that extracts multidimensional concepts by clustering latent difference vectors computed from factual–counterfactual image pairs. These difference vectors, derived from VAE encodings of factual images and their corresponding diffusion-based counterfactuals, capture the classifier-induced transformation in latent

*Corresponding author.
Email addresses: payal.varshney@dfki.de (Payal Varshney), adriano.lucieri@dfki.de (Adriano Lucieri), christoph.balada@dfki.de (Christoph Balada), andreas.dengel@dfki.de (Andreas Dengel), sheraz.ahmed@dfki.de (Sheraz Ahmed)

space. In contrast to CDCT, which modifies individual latent dimensions, CDLC leverages directional clustering to reveal coordinated latent changes that correspond to semantically meaningful concepts. This approach not only reduces computational complexity by eliminating exhaustive per-dimension search but also enables the discovery of classifier-relevant concept directions that emerge from interactions among multiple latent dimensions, effectively overcoming a limitation of the original CDCT formulation.

This paper makes the following key contributions:

- A novel framework, Concept Directions via Latent Clustering (CDLC), is introduced as an extension of CDCT [15], significantly reducing computational complexity compared to dimension-wise search.
- CDLC extracts global, multidimensional semantic concept directions by clustering latent-difference vectors obtained from VAE encodings of factual and diffusion-generated counterfactual images.
- The effectiveness of CDLC is validated on a skin lesion classification task, where the discovered concept directions not only reliably flipped classifier predictions but also transferred robustly to unseen samples.

2. Related Work

Explainable Artificial Intelligence (XAI) aims to make machine learning models more transparent by providing human-understandable insights into their decision-making processes [1, 2]. Among various XAI approaches, concept-based explanations have gained increasing attention due to their ability to link internal model representations to semantically meaningful, human-interpretable concepts [3, 4, 18]. In supervised settings, concept-based methods rely on expert-provided annotations, which are costly and time-consuming to collect [3, 19]. For instance, Testing with Concept Activation Vectors (TCAV) [3] requires example sets for each concept and learns linear concept activation vectors to quantify a model's sensitivity to those concepts. While effective, this reliance on curated concept examples limits the scalability and generalization of such methods. To overcome these limitations, unsupervised concept-based methods have been introduced [4, 20]. A prominent example is Automatic Concept Explanations (ACE) [4], which discovers concepts by clustering segmented image patches, without requiring manual annotation. However, ACE relies on spatial assumptions by considering localized image regions as potential concept candidates, limiting its ability to capture non-local or abstract concepts. A comprehensive survey [18] has reviewed the landscape of concept-based explanations, outlining a range of methodologies and their respective applications.

Another prominent paradigm in XAI is counterfactual explanations, which provide "what-if" scenarios to reveal how minimal semantic changes to an input can alter a model's prediction [21]. With the emergence of generative models, counterfactual synthesis has advanced significantly, generating more realistic and semantically meaningful examples.

In particular, diffusion-based approaches have been proposed to generate counterfactuals by incorporating classifier gradients into the denoising process [22, 23] or by applying adaptive parameterization and cone regularization of gradients [24]. While these methods produce visually realistic counterfactuals, their explanations are typically local to individual samples. In contrast, the Global Counterfactual Directions (GCD) [25] method learns latent directions that consistently invert sample classifications via a proxy model, enabling more generalizable interpretability.

Recent work has attempted to combine concepts with counterfactual generation. CoLa-DCE [14] uses classifier guidance to produce spatially constrained, concept-conditioned counterfactuals. Similarly, DiffEx [26] combines a vision–language model with diffusion semantics to generate a semantic hierarchy of attributes and rank their influence on classifier decisions. However, these methods remain sample-specific and do not yield reusable global concept vectors. In contrast, Varshney et al. [15] leveraged classifier-guided counterfactual trajectories to identify global disentangled semantic concepts. Despite its strengths, CDCT's dimension-wise traversal is computationally expensive and limited in detecting complex, multidimensional concepts.

Another line of research leverages clustering in latent spaces to model semantic variation. For example, DifCluE [27] clusters latent embeddings produced by a diffusion autoencoder to generate diverse counterfactual explanations, effectively modeling intra-class variation, but does not explicitly identify global concept directions. A parallel line clusters concept-attribution vectors rather than latent features. PCX [28] computes per-sample concept-relevance vectors and fits class-wise Gaussian mixture models to cluster them into prototypical decision strategies that explain model behavior.

However, most existing approaches reveal several open challenges: reliance on concept annotations or spatial heuristics, a focus on localized or sample-specific edits, the need for exhaustive latent traversal, and limited capacity to capture global, multidimensional semantic concepts. To address these limitations, a novel framework, Concept Directions via Latent Clustering, is proposed as an extension of CDCT [15]. It discovers global, class-specific concept directions by clustering directional latent differences between factual and counterfactual images. This approach enables the extraction of complex, multidimensional concept directions in an unsupervised manner and offers a computationally efficient framework for concept-based explanation, enhancing both interpretability and scalability.

3. Methodology

Concept Directions via Latent Clustering (CDLC) is a novel framework designed to extract global, multidimensional concept directions from factual and counterfactual image pairs. It builds upon the counterfactual generation stage of CDCT [15]. However, instead of relying on computationally expensive dimension-wise latent traversals, it identifies concept directions

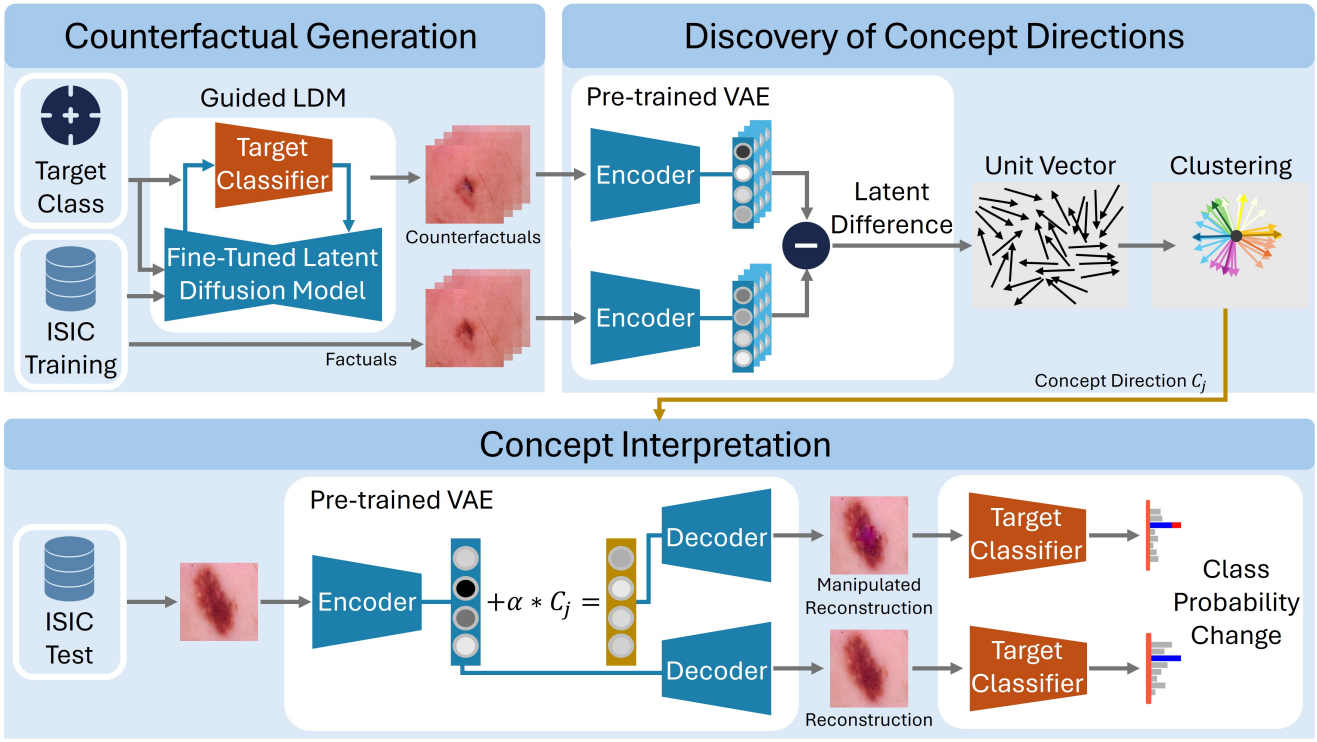


Figure 1: Overview of the CDLC framework. Counterfactuals are generated using a Latent Diffusion Model (LDM) with classifier guidance, following the procedure used in CDCT [15]. Factual–counterfactual image pairs are encoded using a pretrained Variational Autoencoder (VAE), and the difference between their latent representations is normalized to form unit vectors. These vectors are clustered to identify class-specific concept directions C_j. During inference, each direction, scaled by a factor α, is applied to the test sample's latent representation to observe its effect on classifier output.

by clustering unit-normalized difference vectors in a semantic latent space. This section describes the key components of CDLC in detail.

3.1. Counterfactual Generation via Latent Diffusion Models

Given an input image x_f and a trained classifier $f(\cdot)$, a counterfactual image x_{cf} is synthesized to belong to a target class $y_{cf} \neq f(x_f)$. The counterfactual generation process follows the first stage of the CDCT framework [15], termed *Generation of Counterfactual Trajectories*, where classifier guidance is integrated into the denoising steps of a latent diffusion model [16] to produce semantically meaningful modifications aligned with the desired class transition. Details of this process can be found in the CDCT framework [15]. Rather than capturing intermediate reconstructions across the trajectory, only the final counterfactual image is retained. This design choice aims to capture the semantic shift required to alter the classifier's prediction while reducing storage and computational overhead. The resulting counterfactual is visually realistic and deviates minimally from the factual image, thereby isolating class-specific semantic changes.

3.2. Discovery of Concept Directions

Each factual image x_f and its corresponding counterfactual x_{cf} for a specific target class are encoded using a pretrained Variational Autoencoder, resulting in latent representations $\mathbf{z}_f = \text{VAE}(x_f)$ and $\mathbf{z}_{cf} = \text{VAE}(x_{cf})$. These latent embeddings capture high-level semantic characteristics of the original and counterfactual images in a compact form.

A difference vector is computed between the latent representations of each factual and counterfactual image pair as:

$$\Delta\mathbf{z} = \mathbf{z}_{cf} - \mathbf{z}_f$$

Each difference vector is reshaped into a flat vector $\Delta\mathbf{z} \in \mathbb{R}^d$. To focus solely on the directional semantics and eliminate the influence of vector magnitude, each difference vector is normalized to unit length:

$$\tilde{\mathbf{z}} = \frac{\Delta\mathbf{z}}{\|\Delta\mathbf{z}\|_2}$$

This transformation projects the latent differences onto the unit hypersphere in $\mathbb{R}^d$, making them suitable for angular similarity-based clustering.

A collection of such unit-norm latent difference vectors $\{\tilde{\mathbf{z}}_i\}_{i=1}^N$, derived from multiple factual–counterfactual pairs for a given target class, is clustered into K clusters to identify shared semantic transformations. Cluster centers $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ are computed by averaging the unit vectors within each cluster and re-normalizing to unit length:

$$\mathbf{c}_k \in \mathbb{R}^d, \quad \|\mathbf{c}_k\|_2 = 1$$

Each cluster center $\mathbf{c}_k$ represents a global concept direction, a consistent and interpretable shift in the latent space associated with the classifier's decision boundary. These concept directions can then be applied to the latent representations of new

test samples to induce semantically meaningful changes in the reconstructed output.

3.3. Concept Interpretation

To interpret each discovered concept direction $\mathbf{c}_k$, it is applied to the latent encoding of an unseen test sample:

$$\mathbf{z}'_{\text{test}} = \mathbf{z}_{\text{test}} + \alpha \cdot \mathbf{c}_k,$$

where $\alpha \in \mathbb{R}$ controls the strength of the semantic manipulation. The modified latent vector $\mathbf{z}'_{\text{test}}$ is then passed through the VAE decoder to generate a concept-modified image.

In contrast to CDCT, which iteratively modifies each latent dimension to search for influential concepts, CDLC learns global, multidimensional concept directions directly from the difference of factual–counterfactual encodings. CDLC avoids storing intermediate image trajectories and relies on CPU-only clustering of latent differences; no auxiliary VAE training or iterative dimension-wise search is required. These changes lead to a significant increase in computational efficiency and interpretable semantic transformations.

4. Experiments & Results

This section presents the experimental setup and an analysis of the results produced by Concept Directions via Latent Clustering framework. Quantitative and qualitative evaluations assess the semantic coherence of the extracted concept directions and their effectiveness in influencing classifier predictions.

4.1. Dataset and Classification Model

The same dataset and classification architecture used in the CDCT framework [15] are adopted for evaluation. Experiments are conducted on a consolidated dermoscopic image dataset derived from the International Skin Imaging Collaboration (ISIC) challenges (2016–2020)[1]. Following the duplicate removal strategy proposed by Cassidy et al. [29], a curated dataset of 29,468 unique images is obtained and stratified into training, validation, and test subsets. The dataset includes eight diagnostic categories: *Melanocytic Nevus* (NV), *Melanoma* (MEL), *Basal Cell Carcinoma* (BCC), *Actinic Keratosis* (AK), *Benign Keratosis* (BKL), *Dermatofibroma* (DF), *Vascular Lesions* (VASC), and *Squamous Cell Carcinoma* (SCC).

A ResNet-50 [30] model is trained on the ISIC training partition and used as the target classifier throughout all experiments. This network is used to generate classifier-guided counterfactuals and evaluate the effectiveness of discovered concept directions.

4.2. Generation of Counterfactuals

The same classifier-guided counterfactual generation approach introduced in the CDCT framework [15] is adopted, with a simplified setting. Rather than capturing intermediate reconstructions across the trajectory, only the final counterfactual image is retained. Counterfactuals are generated using a Latent Diffusion Model based on the Stable Diffusion (SD) 2.1 architecture[2], which is fine-tuned on the consolidated ISIC training dataset to align the generative manifold with the dermoscopic image domain. For counterfactual generation, the same hyperparameters (guidance scale, diffusion steps) are adopted as reported in CDCT, to ensure consistency and comparability (see Appendix A for hyperparameters). For each training sample, counterfactuals are synthesized for all target classes other than the class predicted by the classifier.

4.3. Extraction of Concept Directions

Semantic concept directions induced by classifier-guided counterfactuals are explored in two distinct latent spaces: (1) the pretrained encoder of the Stable Diffusion 2.1 model, referred to as the *LDM encoder*, and (2) a Variational Autoencoder trained on counterfactual trajectories, referred to as the *CDCT encoder* (see Step 2 of the CDCT framework [15] for further details). Experimental setup, results, and analysis based on the CDCT encoder are provided in the Supplementary Material (Appendix E, F).

To enable class-specific concept discovery, latent difference vectors are computed separately for each target class. For this purpose, training samples not predicted as the target class are selected, and counterfactual images are generated that shift the classifier's prediction to the target class. The corresponding factual and counterfactual images are encoded using the LDM encoder, resulting in latent embeddings of shape $4 \times 32 \times 32$. The element-wise difference between these embeddings captures the transformation required to alter the prediction. Each difference tensor is flattened into a 4096-dimensional vector and normalized to unit length, forming a directional vector in latent space.

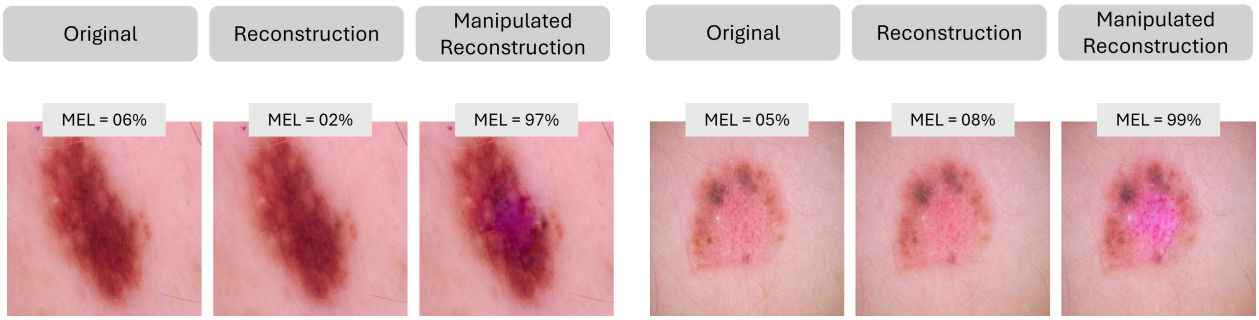
The unit vectors are collected for all training samples that are not predicted as a given target class. Spherical K-Means clustering [31] is then applied to group these vectors that reflect similar semantic transformations. The number of clusters K is selected based on the highest silhouette score (see Appendix A for per-class K values). A detailed ablation study analyzing the effect of the number of clusters K is provided in the Supplementary Material (Appendix B). Each resulting cluster represents a set of consistent latent transformations, and the average unit direction within a cluster is interpreted as a representative concept direction for the target class.

4.4. Interpretation and Evaluation of Concept Directions

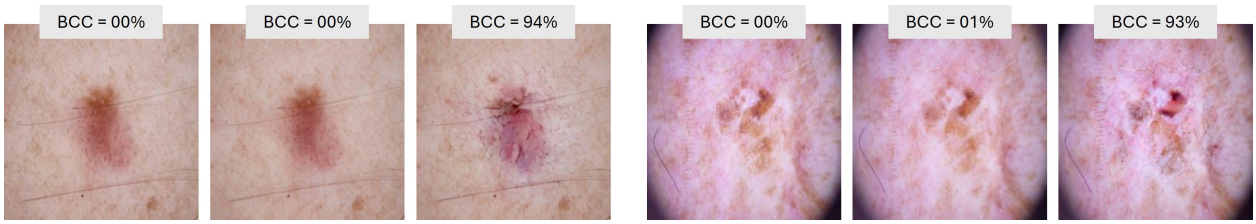
To assess the interpretability and effectiveness of the extracted concept directions, each identified concept direction is

[1] Dataset available at: `https://challenge.isic-archive.com/challenges/`.

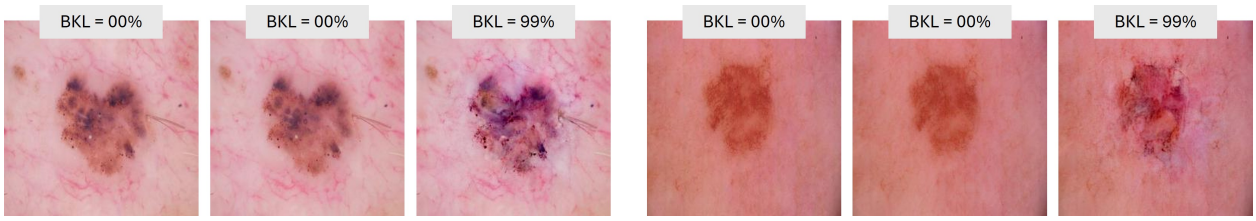
[2] Model available at: `https://huggingface.co/stabilityai/stable-diffusion-2-1`.



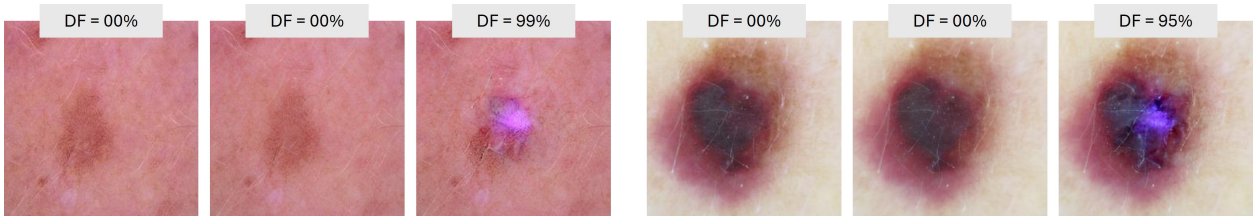
(a) Success Rate = 70.54 % - Purplish Core Pigmentation.



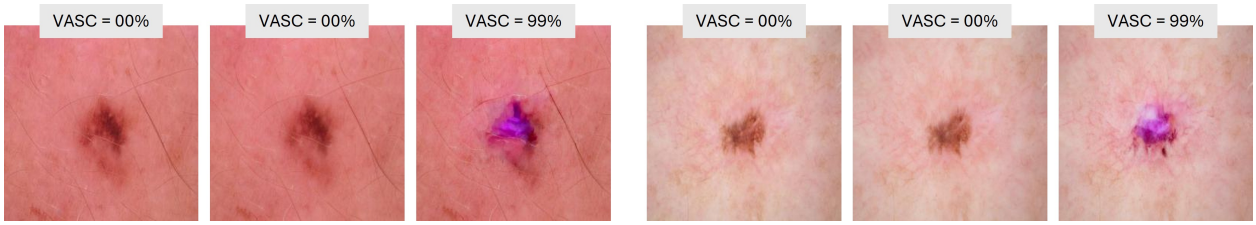
(b) Success Rate = 83.61 % - Reddish Core with Blue-Gray Dots.



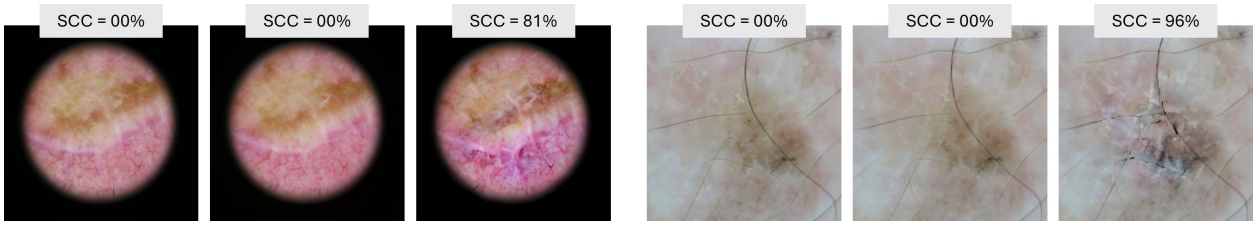
(c) Success Rate = 81.09 % - Blotchy Pigmentation with Irregular Texture.



(d) Success Rate = 85.44 % - Central Pinkish Veil with Asymmetric Intensified Pigmentation.



(e) Success Rate = 84.59 % - Central Purplish Veil.



(f) Success Rate = 70.47 % - White Structures with Irregular Vessels.

Figure 2: Discovered concepts by CDLC on the ISIC dataset using the LDM encoder. Each row shows two examples: original, reconstructed, and manipulated reconstruction (left to right). The predicted probability for the target class associated with each concept direction is shown above each image.

applied to unseen test images in the latent space. A detailed ablation study analyzing the effect of the scaling parameter α on concept traversal is provided in the Supplementary Material (Appendix C). Figure 2 illustrates concept directions identified by CDLC using the LDM encoder. For *Melanoma*, the extracted direction (Figure 2a) increases central pigmentation and introduces a purplish hue in the core lesion. In *Basal Cell Carcinoma*, detected concept direction (Figure 2b) produces a reddish-purple hue with blue-gray dots. For *Benign Keratosis*, a concept direction (Figure 2c) induces purplish blotches, irregular pigmentation, and blurred borders. In the case of *Dermatofibroma*, the direction (Figure 2d) introduces a violet-toned center with subtle border irregularities and increased asymmetric pigmentation. For *Vascular Lesions*, the direction (Figure 2e) enhances central coloration with purplish tones and peripheral pink diffusion. Finally, *Squamous Cell Carcinoma* directions (Figure 2f) reveal white central structures accompanied by irregular linear streaks suggestive of vascular features.

Table 1 reports the quantitative results for the detected concept directions. Success Rate (SR) measures the proportion of test samples where traversal along the direction increases the probability of the target class, relative to the reconstructed image. LPIPS [32] and FID [33] assess perceptual and distributional realism, while TCAV quantifies concept alignment with the model's decision boundary, computed from the final conv3 outputs of the third bottleneck block in layer 4 of ResNet50. A detailed analysis of TCAV computed across different layers of ResNet50 for these concept directions can be found in the Supplementary Material (Appendix D). Most concepts achieve high SR (70–85%), confirming their strong influence on classifier decisions. Concepts such as *Central Pinkish Veil* and *Central Pinkish Veil with Asymmetric Intensified Pigmentation* attain both high SR and near-perfect TCAV, indicating consistent use by the model. In contrast, *White Structures with Irregular Vessels* show lower SR and TCAV, suggesting weaker alignment. Overall, SR, LPIPS, FID, and TCAV demonstrate that the discovered concepts are semantically meaningful and predictive of model behavior.

Table 1: Quantitative results for the detected concept directions, reported in terms of Success Rate (SR), LPIPS, FID, and TCAV. Together, these metrics demonstrate that the discovered concepts are both semantically meaningful and predictive of model behavior.

Concept	**SR (%) ↑**	**LPIPS ↓**	**FID ↓**	**TCAV ↑**
Purplish Core Pigmentation	70.5	0.12	30.5	0.97
Reddish Core with Blue-Gray Dots	83.6	0.17	43.1	0.82
Blotchy Pigmentation with Irregular Texture	81.1	0.20	47.0	0.92
Central Pinkish Veil with Asymmetric Intensified Pigmentation	85.4	0.15	47.1	1.00
Central Purplish Veil	84.6	0.15	53.5	1.00
White Structures with Irregular Vessels	70.5	0.18	51.2	0.64

4.5. Computational Efficiency

To assess computational efficiency, CDLC is compared with CDCT for detecting concept directions of the *Melanoma* class. As shown in Table 2, both methods have comparable generation times, but CDLC requires ∼4.6× less storage by retaining only latent differences instead of entire trajectories. Beyond storage, CDCT requires additional architectural overhead, including training an auxiliary VAE and performing iterative dimension-wise traversal. In contrast, CDLC requires no additional training; concept discovery reduces to clustering latent differences, which runs efficiently on CPU. Despite running on CPU, CDLC represents a ∼5.3× speedup over CDCT's GPU-based extraction step. Overall, CDLC substantially reduces computation time and storage, offering a scalable and practical alternative without sacrificing interpretability.

Table 2: Runtime and storage comparison between CDLC and CDCT for detecting concept directions of the *Melanoma* class. All GPU operations were performed on an NVIDIA L40S GPU. CDLC achieves ∼4.6× lower storage requirements and ∼5.3× faster concept extraction despite running on CPU.

Method	**Hardware**	**Time**	**Storage**
CDCT (trajectory)	GPU	1d 16h 25m	314 MB
CDLC (embed. diff.)	GPU	1d 18h 35m	69 MB
CDCT (dim. search)	GPU	∼80 m	—
CDLC (clustering)	CPU	∼15 m	—

5. Discussion

The results demonstrate that CDLC effectively extracts global concept directions for skin lesion classification while significantly improving computational efficiency over CDCT. By avoiding iterative traversal of individual latent dimensions, CDLC reduces the overhead typically associated with concept discovery, especially in high-dimensional latent spaces. CDLC achieves a ∼4.6× reduction in storage requirements and a ∼5.3× speedup in concept extraction relative to CDCT, and clustering in CDLC executes entirely on CPU without requiring GPU resources. Additionally, it captures rich, multidimensional semantic directions beyond the axis-aligned perturbations of CDCT, which tend to detect only simple features. Beyond efficiency, CDLC reliably discovers class-specific concept directions by constructing latent difference vectors separately for each target class. This change avoids feature entanglement observed in CDCT, where a single VAE is trained on trajectories across all classes. As a result, the extracted directions are more semantically coherent and clinically interpretable.

To further analyze the visual characteristics of the extracted directions, the behavior of the two latent encoders used in the framework is compared. While both encoders support meaningful concept discovery, the CDCT encoder often fails to reconstruct subtle yet clinically relevant features, limiting the fidelity of concept manipulations. In contrast, the LDM encoder preserves fine-grained details and structural consistency, yielding more realistic and diagnostically coherent manipulations. Nonetheless, directions in the LDM space frequently

introduce pinkish or violet hues, which may influence the visual appearance of concepts. Possible explanations include the fact that CDLC learns directions in latent difference space and does not explicitly optimize for hue variations; such hue shifts may therefore arise from generative priors in the latent-diffusion model, from color correlations learned by the target classifier, or from relevant color biomarkers. Therefore, color changes alone are not interpreted as concepts; instead, the qualitative analysis emphasizes structural and textural cues that consistently co-occur with the discovered directions.

The concept directions discovered by CDLC demonstrate alignment with established dermoscopic features. For instance, directions targeting *Melanoma* often exhibit asymmetric pigmentation and color variegation, features of malignancy reported in clinical literature [34]. In *Basal Cell Carcinoma*, the extracted directions introduce translucent reddish or violet hues with dark pigment patches, consistent with features such as ulceration and blue-gray dots [35]. For *Dermatofibroma*, directions emphasize a central scar-like area of pink hues reported in literature [36], while *Vascular Lesions* display purplish centers with surrounding pink diffusion, resembling the vascular blush of angiomas [37]. Similarly, directions for *Squamous Cell Carcinoma* reveal white scaly textures and irregular vascular patterns, characteristic of keratinizing and sun-damaged lesions [38].

In addition to replicating known diagnostic cues, CDLC also uncovers subtler variations such as peripheral pigmentation or central hue changes, which may indicate dataset-specific biases or underexplored clinical markers, and requires clinical feedback. In contrast, reliable concept directions could not be identified for the *Nevus* class. This is likely due to class imbalance, as the dominance of *Nevus* in the training data results in fewer samples being mapped as counterfactual to it. Consequently, the limited number of latent difference vectors hinders the discovery of consistent semantic directions for this class.

6. Limitations & Future work

While CDLC demonstrates promising results in extracting semantically meaningful and class-specific concept directions, several limitations remain. The reliability of extracted directions depends on the quality of counterfactuals, as artifacts or semantic drift at this stage can introduce noise into the latent differences. Some concept directions may exhibit hue shifts (e.g., pink/violet), potentially reflecting generative priors, color correlations in the classifier, or potential color biomarkers. Disambiguation of these causes would require targeted controls (e.g., color-invariant rendering, grayscale analyses) and clinical validation, which we consider beyond the scope of this work and leave for future work. Additionally, spherical clustering assumes that concept directions are well-separated based on angular similarity on a unit hypersphere, which may not adequately capture complex latent geometries or overlapping semantic structures. Exploring more flexible clustering approaches (e.g., kernelized or manifold-based methods) remains an avenue for future work.

Future work could explore alternative encoders, such as Diffusion Autoencoders [39], to enhance reconstruction fidelity and semantic expressiveness. Extending CDLC to other modalities, validating its transferability across domains, and testing it with alternative architectures may further broaden its applicability. Finally, incorporating human-in-the-loop evaluation or clinical feedback would be essential to establish the medical validity of the discovered concepts and support their real-world integration. Such validation requires substantial time and collaboration with domain specialists and is left for future work.

7. Conclusion

This work introduced CDLC, a framework for discovering global, class-specific concept directions by clustering latent differences between factual and counterfactual image pairs. Experiments on a skin lesion classification task demonstrated that the extracted directions influence model predictions and align with known dermoscopic features, supporting their interpretability. However, some concept directions, especially obtained from the LDM encoder, consistently exhibit pinkish or purplish hues, irrespective of the target class. This visual redundancy may indicate an inherent bias in the generative prior, suggesting a need for further analysis to disentangle meaningful concepts from model or dataset-specific artifacts. The ability to uncover diverse, clinically relevant concepts positions CDLC as a valuable tool for advancing concept-based explainability in high-stakes domains.

8. Acknowledgments

The project was funded by the Federal Ministry for Education and Research (BMBF) with grant number 03ZU1202JA.

9. CRediT authorship contribution statement

Payal Varshney: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Adriano Lucieri:** Conceptualization, Data curation, Supervision, Validation, Writing – review and editing. **Christoph Balada:** Writing – review and editing. **Andreas Dengel:** Funding acquisition, Project administration, Resources, Supervision, Writing – review and editing. **Sheraz Ahmed:** Supervision, Writing – review and editing.

10. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-4o in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[2] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[3] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.

[4] A. Ghorbani, J. Wexler, J. Y. Zou, B. Kim, Towards automatic concept-based explanations, Advances in neural information processing systems 32 (2019).

[5] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[7] D. Shin, The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai, International journal of human-computer studies 146 (2021) 102551.

[8] W. Guo, Explainable artificial intelligence for 6g: Improving trust between human and machine, IEEE Communications Magazine 58 (6) (2020) 39–45.

[9] D. Das, S. Banerjee, S. Chernova, Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery, in: Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction, 2021, pp. 351–360.

[10] M. Ebers, Regulating explainable ai in the european union. an overview of the current legal framework (s), An Overview of the Current Legal Framework (s)(August 9, 2021). Liane Colonna/Stanley Greenstein (eds.), Nordic Yearbook of Law and Informatics (2020).

[11] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, et al., The role of explainable ai in the context of the ai act, in: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 2023, pp. 1139–1150.

[12] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, et al., Explaining in style: training a gan to explain a classifier in stylespace, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 693–702.

[13] M. Atad, V. Dmytrenko, Y. Li, X. Zhang, M. Keicher, J. Kirschke, B. Wiestler, A. Khakzar, N. Navab, Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan, arXiv preprint arXiv:2207.07553 (2022).

[14] F. Motzkus, C. Hellert, U. Schmid, Cola-dce–concept-guided latent diffusion counterfactual explanations, arXiv preprint arXiv:2406.01649 (2024).

[15] P. Varshney, A. Lucieri, C. Balada, A. Dengel, S. Ahmed, Generating counterfactual trajectories with latent diffusion models for concept discovery, in: International Conference on Pattern Recognition, Springer, 2025, pp. 138–153.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[17] D. P. Kingma, M. Welling, et al., Auto-encoding variational bayes (2013).

[18] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, E. Baralis, Concept-based explainable artificial intelligence: A survey, arXiv preprint arXiv:2312.12936 (2023).

[19] A. Lucieri, M. N. Bajwa, A. Dengel, S. Ahmed, Explaining ai-based decision support systems using concept localization maps, in: International Conference on Neural Information Processing, Springer, 2020, pp. 185–193.

[20] T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, T. Serre, Craft: Concept recursive activation factorization for explainability, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2711–2721.

[21] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[22] G. Jeanneret, L. Simon, F. Jurie, Diffusion models for counterfactual explanations, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 858–876.

[23] G. Jeanneret, L. Simon, F. Jurie, Adversarial counterfactual visual explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16425–16435.

[24] M. Augustin, V. Boreiko, F. Croce, M. Hein, Diffusion visual counterfactual explanations, Advances in Neural Information Processing Systems 35 (2022) 364–377.

[25] B. Sobieski, P. Biecek, Global counterfactual directions, in: European Conference on Computer Vision, Springer, 2024, pp. 72–90.

[26] T. Kazimi, R. Allada, P. Yanardag, Explaining in diffusion: Explaining a classifier with diffusion semantics, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 14799–14809.

[27] S. Jain, A. Sangroya, L. Vig, Difclue: Generating counterfactual explanations with diffusion autoencoders and modal clustering, arXiv preprint arXiv:2502.11509 (2025).

[28] M. Dreyer, R. Achtibat, W. Samek, S. Lapuschkin, Understanding the (extra-) ordinary: Validating deep model decisions with prototypical concept-based explanations, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2024, pp. 3491–3501.

[29] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, M. H. Yap, Analysis of the isic image datasets: Usage, benchmarks and recommendations, Medical image analysis 75 (2022) 102305.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[31] K. Hornik, I. Feinerer, M. Kober, C. Buchta, Spherical k-means clustering, Journal of statistical software 50 (2012) 1–22.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).

[34] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, et al., Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet, Journal of the American Academy of Dermatology 48 (5) (2003) 679–693.

[35] D. Altamura, S. W. Menzies, G. Argenziano, I. Zalaudek, H. P. Soyer, F. Sera, M. Avramidis, K. DeAmbrosis, M. C. Fargnoli, K. Peris, Dermatoscopy of basal cell carcinoma: morphologic variability of global and local features and accuracy of diagnosis, Journal of the American Academy of Dermatology 62 (1) (2010) 67–75.

[36] A. L. C. Agero, S. Taliercio, S. W. Dusza, C. Salaro, P. Chu, A. A. Marghoob, Conventional and polarized dermoscopy features of dermatofibroma, Archives of dermatology 142 (11) (2006) 1431–1437.

[37] V. Piccolo, T. Russo, E. Moscarella, G. Brancaccio, R. Alfano, G. Argenziano, et al., Dermatoscopy of vascular lesions, Dermatol Clin 36 (4) (2018) 389–395.

[38] C. Rosendahl, A. Cameron, G. Argenziano, I. Zalaudek, P. Tschandl, H. Kittler, Dermoscopy of squamous cell carcinoma and keratoacanthoma, Archives of dermatology 148 (12) (2012) 1386–1392.

[39] K. Preechakul, N. Chatthee, S. Wizadwongsa, S. Suwajanakorn, Diffusion autoencoders: Toward a meaningful and decodable representation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10619–10629.