Interactive Visual Exploration of Latent Spaces for Explainable AI: Bridging Concepts and Features*

Carlo Metta[1,*], Eleonora Cappuccio[1,2,3] and Salvatore Rinzivillo[1]

[1]CNR-ISTI
[2]Università di Pisa, Pisa, Italy
[3]Università degli studi di Bari Aldo Moro, Bari, Italy

Abstract

Latent space exploration offers a powerful lens for interpreting and improving the explainability of black-box AI models. This paper introduces a visual interface based on β-Variational Autoencoders that enables users to navigate latent spaces interactively. By employing tools like visual latent sliders and transformation pathways, the interface demonstrates how latent dimensions can influence image representations, uncovering semantic structure and disentangled features. Using the **MedMNIST** medical image dataset, we illustrate the potential of this approach to bridge the gap between technical latent space analysis and intuitive understanding. Although the focus is on presenting the methodology, this work sets the stage for integrating user interaction and metrics, particularly in high-stakes domains such as medical imaging.

Keywords

XAI, Latent Space, Healthcare, Interactive Interfaces

1. Introduction

The rise of black-box machine learning models, particularly deep neural networks, has revolutionized numerous domains, including healthcare [1, 2]. However, their opaque nature raises concerns about trust, accountability, and decision reliability, particularly in high-stakes applications like medical diagnosis and treatment planning. Explainable Artificial Intelligence (XAI) seeks to address these challenges by providing mechanisms to interpret and understand model decisions [3, 4]. Within XAI, methods focusing on latent space exploration have emerged as a promising avenue to bridge the gap between complex model representations and human interpretability [5, 6].

Latent space, the compressed representation of data learned by models such as autoencoders and generative networks, encodes rich semantic structures that can provide insight into the underlying data distributions and decision pathways of models [7, 8]. In particular, β-Variational Autoencoders (β-VAEs) have demonstrated the ability to disentangle latent factors, offering interpretable dimensions that align with meaningful attributes [6]. Despite these advances, effectively exploiting the latent space for interpretability remains a challenge, especially in healthcare, where actionable insights are critical [9].

Healthcare presents unique challenges and opportunities for XAI. Models trained on medical data must provide not only accurate predictions, but also explanations that resonate with medical experts [10]. Techniques such as counterfactual reasoning and prototype generation have shown promise in this regard [11, 12], yet their potential is largely unexplored in the context of latent space navigation, with some initial efforts beginning to address this challenge [13, 14]. A human-centered approach, which incorporates user-driven exploration and visualization, can enhance the interpretability and usability of AI systems [15, 16].

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
carlo.metta@isti.cnr.it (C. Metta); eleonora.cappuccio@phd.unipi.it (E. Cappuccio); salvatore.rinzivillo@isti.cnr.it (S. Rinzivillo)
0000-0002-9325-8232 (C. Metta); 0000-0002-6105-2512 (E. Cappuccio); 0000-0003-4404-4147 (S. Rinzivillo)

In this work, we propose an interactive visual interface for latent space exploration and interpretation, designed to complement existing XAI methods. Centered on β-VAEs, our work introduces an interactive tool, based on grid visualization of latent pathways, to enable intuitive navigation of latent spaces. Using this tool, users can identify significant latent dimensions, uncover semantic structures, and explore counterfactuals. While our initial focus is methodological, we propose this visual interface as a step toward human-centered AI applications in high-stakes domains like healthcare.

The remainder of this paper is organized as follows. Section 2 summarizes related work in XAI, latent space exploration, and healthcare applications. Section 3 presents the proposed interface and its key components. Section 4 discusses qualitative results using common benchmark datasets and the **MedMNIST** dataset, and Section 5 concludes with a discussion of future directions.

2. Related Work

Explainable AI (XAI) has become a critical research area that aims to address the opacity of black-box machine learning models. Early work focused on post hoc methods, such as saliency maps [17] and feature importance scores [10], which provide local explanations for individual predictions. Although these methods have been widely adopted, they often lack the semantic clarity needed for domain experts, particularly in high-stakes applications such as healthcare care [9]. Recent advances emphasize interactive and human-centered approaches to explainability, aligning model outputs with user mental models [15].

Latent space representations, a cornerstone of representation learning, have opened new pathways for interpretability. Techniques such as autoencoders [8, 7] and β-variational autoencoders (β-VAE) [6] enable compact and structured representations of high-dimensional data. These models have demonstrated the potential to disentangle latent factors into interpretable dimensions, allowing users to explore data semantics intuitively. Visualization techniques, such as t-SNE and UMAP, have further facilitated latent space interpretation by projecting high-dimensional embeddings in lower dimensions for human analysis [18, 19].

Counterfactual reasoning has emerged as a powerful tool within XAI, providing explanations by presenting "what-if" scenarios that clarify model decisions [11, 20]. This approach has been particularly impactful in the healthcare domain [21]. From a Human-centered point of view, counterfactuals are considered to be more intuitive and less cognitive demanding, as they align with the ways in which humans produce explanations [22]. These methods often leverage latent space representations to generate plausible counterfactuals that align with the data distribution. For instance, models like ABELE [23] generate counterexamples in the latent space to explain classification outcomes. However, these approaches are limited in their interactivity and do not fully leverage the exploratory potential of latent dimensions.

Despite these advancements, there is a notable gap in integrating latent space navigation into interactive environments that supports explainability. While some initial efforts have explored interactive latent space visualization, such as tools for multimodal data exploration and annotation [24, 25], these works primarily focus on technical demonstrations and general-purpose applications. Furthermore, existing methods often overlook the human-centered design principles necessary for effective deployment in critical domains like healthcare, producing examples that are impractical or inconsistent within the domain [26, 21].

This work aims to address these gaps by proposing an interface that combines β-VAEs with interactive tools such as visual latent sliders and transformation pathways. Our approach emphasizes the interpretability of latent dimensions and their semantic contributions to model decisions, bridging the divide between technical latent space analysis and practical user needs. By complementing existing methods and integrating human-centered design principles, we provide a foundation for explainability models that are both robust and user-friendly.

3. Methodology

3.1. Overview

The latent space of an autoencoder is a high-dimensional representation that encodes the structure and semantics of the input data. However, these latent dimensions are inherently semantically opaque, making their interpretation and use for explainability challenging. In this work, we propose an interactive interface that empowers users to navigate and interpret the latent space through visual exploration and intuitive interactions.

The key idea is to allow users to navigate through the latent dimensions using a visual tool and observe the corresponding changes in the generated images. The latent space is discretized into a grid of regular points. The granularity of the grid can be adjusted to obtain the desired level of detail.

This enables users to discover the semantic properties encoded in the latent features without requiring explicit labelling or understanding of these dimensions. The process is depicted in Figure 2. To illustrate the potential of this approach, we introduce the "Identikit Game," an engaging visualization strategy that allows users to iteratively transform an image into a target by exploring the latent space.

3.2. Latent Space Navigation and Visualization

To facilitate the exploration of latent space, we developed a visual interface where users can interact with each latent dimension through a grid of images. Each line of the grid adjusts a specific latent feature, and the corresponding changes in the generated image are displayed in real time. By clicking on an image, users can intuitively explore the relationships between latent dimensions and their semantic effects on the target.

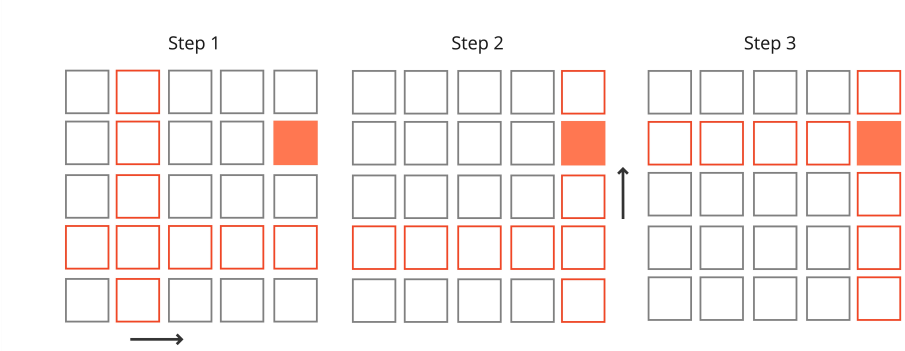

Figure 1: From an initial image, the user can get to the target image, in orange, by moving along the grid.

We employ disentangled representations learned by β-VAEs to ensure that each latent dimension captures an independent feature, minimizing overlap between factors. This disentanglement is critical for enabling meaningful interactions, as changes in one dimension do not inadvertently affect others.

3.3. The Identikit Game: An Interactive Approach

The "Identikit Game" is a user-centric approach designed to make latent space exploration both intuitive and engaging. Users are presented with a starting image X and a target image Y. The objective is to iteratively adjust the latent dimensions to transform X into Y by exploring the latent space.

The Identikit Game is designed to enable users to dynamically and interactively explore latent spaces by navigating through its coordinate axes. In this game, the user begins with an initial image, which is represented by a latent vector in a high-dimensional latent space. This latent space is typically structured such that each axis represents a specific latent dimension learned by the model, each encoding a distinct feature or property of the input data.

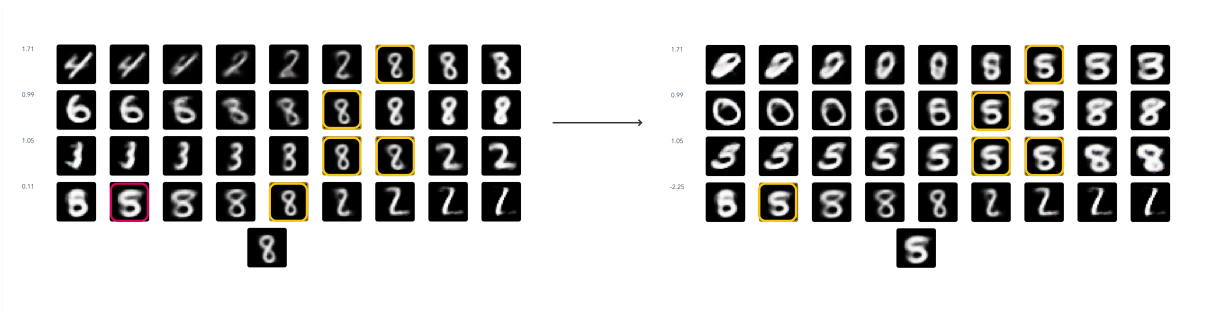

Figure 2: An example of the interface with the **MNIST** dataset. On the bottom, the single image shows the reconstructed element of the point resulting from the grid selection. The grid on the top shows the reconstructed images of the regular points of the latent space. In this example, the user can navigate from the number 8 to the number 5 by selecting first an image as close as possible to the target one, highlighted in the red box, that resembles the number 5. In the second step, the number of candidate images for number 5 is larger, and the user can choose those that are closer to the internal representation.

The user interacts with the game by selecting a starting point in this latent space, corresponding to a specific latent vector that generates the initial image. From this point, the user can move along any of the axes in the latent space, making discrete steps toward a new point. As the user adjusts the latent vector along these axes, the corresponding image is updated in real-time, providing immediate visual feedback.

The core idea of the game is that the user explores different latent dimensions and sees in advance how changes in those dimensions affect the generated image. For example, if the user selects an axis associated with the ”smiling” feature, they will observe how moving along that axis influences the smile of a face in the image. The user can choose to follow any latent axis that interests them and adjust the vector step by step, observing how each movement transforms the image.

At each step, the game allows the user to see a preview of the reconstructed image they are about to reach by adjusting the latent vector along the selected axis. This step-wise movement makes it easy for the user to explore different regions of the latent space without being overwhelmed by the high dimensionality. They can iterate through multiple dimensions and move from one point to another, following various axes as they try to understand the structure and properties of the latent space.

The interactive nature of the game provides a rich, engaging experience where the user gradually fine-tunes the image, refining their exploration of the latent space. The ultimate goal is to transform the initial image into a target image by manipulating the latent space dimensions, step by step. Users can continue to adjust their movement, following different paths along the axes, until they are satisfied with the generated image.

This iterative process not only helps users understand the latent structure encoded in the model but also encourages a deep connection between the user’s intuitive understanding of features and the underlying latent space representation. Through this hands-on interaction, users can uncover complex, abstract patterns within the data, exploring how the model encodes and decodes meaningful features dynamically.

3.4. Human-Centered AI: Concepts over Features

In the context of explainable AI, an important distinction lies between semantic concepts that are explicitly defined –such as colors, shapes, or sizes– and the concepts that are deeply rooted in the human mind, which emerge from lived experiences and cognitive processes. Explicit semantic concepts are generally well-defined and universally agreed upon, allowing for clear boundaries and easily interpretable representations. For example, the color red, the shape of a circle, or the size of an object can be classified in a consistent and measurable way, making them suitable for many traditional machine-learning tasks that rely on supervised learning. These concepts can be directly mapped to distinct features and labels, making them ideal for applications where clear definitions are necessary.

On the other hand, concepts planted in the human mind are often more abstract and context-dependent, arising from an individual's unique experiences, cultural background, and sensory interactions with the world. These concepts are less tangible and harder to define, as they often overlap with emotions, memories, or social constructs. For instance, the concept of beauty or danger is not universally defined and can be influenced by a variety of factors, including personal experiences, societal norms, and emotional responses. In high-stakes scenarios like healthcare, where concepts such as the severity of disease or treatment efficacy depend on the clinician's experience and the quality of the available data, relying on implicit, individual interpretations of latent features can empower clinicians to make more informed decisions without being constrained by predefined labels or categories. This approach encourages a more nuanced understanding, where the clinician can interpret complex data representations based on their expertise and situational context, rather than being confined to rigid, pre-determined categories.

This distinction becomes crucial in our work, which aims to allow for the exploration of latent spaces without explicitly defining or categorizing these concepts. By leaving the visualization of latent features to the human mind, we avoid the complex task of disentangling and assigning labels to concepts that might be ambiguous or too subjective. Unlike supervised machine learning methods that require clear definitions and annotations, our approach encourages the user to navigate the latent space in a way that aligns with their individual understanding. Users rely on their capacity to intuitively grasp complex, context-dependent features without needing pre-existing labels. This method allows us to sidestep the challenges associated with defining and labeling latent dimensions that may represent abstract or multifaceted concepts, which are often difficult to separate or interpret within a fixed framework.

A core strength of this methodology lies in its alignment with human-centered AI principles, which mirror human reasoning [27, 28]. Latent features, while semantically opaque, can be understood by humans through interaction and observation. By engaging with the latent space, users can detect patterns, correlations, and features that are meaningful to them without needing explicit labels or predefined categories. For instance, a user exploring the latent space of facial images might identify a latent dimension associated with "smiling" simply by observing the transformation of images on the grid. The key is that users work with concepts—abstract, human-recognizable patterns—rather than technical features, making the exploration accessible and intuitive. Interactive methods further enhance this process by allowing users to iteratively refine their understanding of patterns in the latent space, reinforcing the intuitive connection to human-recognizable concepts [29].

In this way, the methodology empowers users to discover meaningful semantic structures, enabling them to interpret latent representations in a flexible, dynamic manner that is aligned with their cognitive processes and expertise. Unlike traditional models that impose predefined semantic frameworks, our approach fosters a more personalized interaction with the latent space, making the system more adaptable and interpretable for diverse users, including clinicians in healthcare or experts in other high-stakes fields.

3.5. Applications in Explainable AI

Our tool has several applications in the field of Explainable AI (XAI): Counterfactual explanations enable users to generate counterfactual examples by manipulating latent dimensions, allowing them to observe how minimal changes in the latent space lead to significant changes in the output. For example, in a medical context, this might involve identifying the smallest change in latent space required to shift a diagnosis from "benign" to "malignant." Similarly, exploring clusters in the latent space facilitates prototype generation, where users can identify representative examples for specific classes or categories. Moreover, interactive exploration helps uncover biases in the latent space, such as dimensions encoding socially sensitive attributes, thereby enabling corrective actions to be taken.

3.6. Benefits of Interactive Exploration

To create a more effective human-centred solution, it is essential to incorporate interactivity into the design of explanation interfaces, as users' interactions extend beyond simply receiving an XAI output and persist until a meaningful understanding is reached [30, 31, 32]. Interactive latent space navigation provides several advantages over traditional XAI methods: By involving users in the exploration process, the methodology fosters a deeper understanding of model behavior[33]. Users can leverage their personal pattern recognition abilities and intuition to make sense of complex latent spaces without requiring technical expertise[29]. Eventually, The visual interface can be adapted to various datasets and use cases, making it a versatile tool for XAI.

4. Experiments

4.1. Overview of Experiments

To evaluate the proposed tool, we conducted a series of experiments across multiple datasets, including **MNIST**, **Fashion MNIST**, **EMNIST**, and **MedMNIST**. The experiments were designed to assess the tool's ability to support latent space exploration and improve explainability. Specifically, we investigated the impact of different latent dimensions on usability, applied dimensionality reduction techniques for visualization, and proposed metrics to determine the optimal latent dimension size for human interaction.

4.2. Datasets and Experimental Setup

The datasets used in our experiments were selected to represent a wide range of data structures and challenges, ensuring that our tool could handle different types of input data and provide meaningful interpretations across various domains:

- **MNIST and EMNIST**: These datasets contain grayscale images of handwritten digits and letters, respectively. MNIST is widely used in machine learning research and consists of 28x28 pixel images of digits from 0 to 9. EMNIST extends MNIST to include letters from the English alphabet, providing a larger set of classes with more varied intra-class variability. These datasets offer relatively straightforward examples of image classification, making them ideal for testing basic functionality of the tool.
- **Fashion MNIST**: This dataset consists of grayscale images of clothing items, such as t-shirts, dresses, and shoes, in 28x28 pixel format. Fashion MNIST is more complex than MNIST due to the high intra-class variability of the clothing items. This dataset was used to test how well the tool can handle images with more diverse features and how the user can explore relationships between latent dimensions related to clothing categories.
- **MedMNIST**: The MedMNIST dataset consists of medical images, both in grayscale and RGB, relevant for clinical applications. This dataset includes categories such as skin lesions, chest X-rays, and blood cell images. It is much more complex than the other datasets, involving data from high-stakes domains like healthcare, where accurate interpretation of latent space representations is critical. Using MedMNIST allowed us to test how well the tool supports the interpretability of complex, domain-specific data in a high-stakes context.

For each dataset, we trained a β-VAE with varying latent dimensions ($d = 4, 8, 16, 32, 64$), ensuring a balance between complexity and interpretability. For RGB datasets, where ideal latent dimensions are numerous, we implemented a ranking mechanism to select and display only the most semantically significant dimensions.

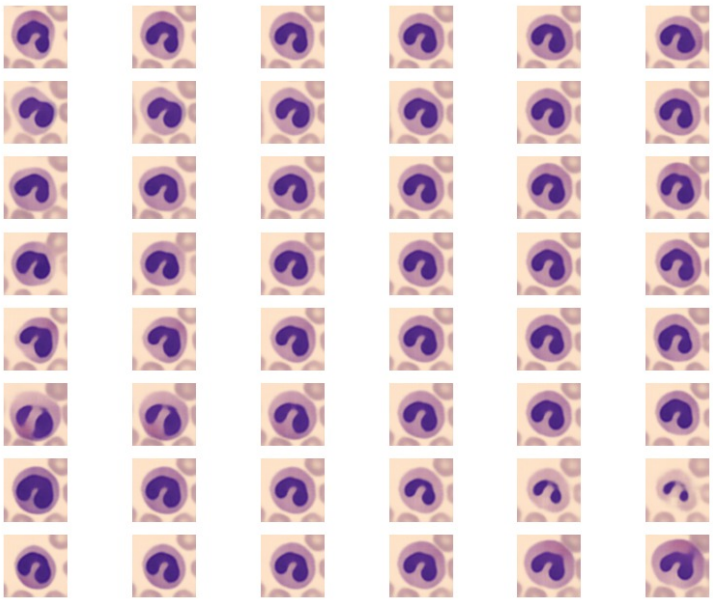

Figure 3: Images from **MedMNIST** dataset of blood cells. Each row displays a sequence showing gradual changes in the appearance of a cell in the latent space.

4.3. Ranking Latent Dimensions

One of the main challenges when working with high-dimensional latent spaces is the cognitive load required to navigate and understand the vast amount of information encoded across multiple dimensions. To reduce this cognitive burden, we focus on selecting only the most informative latent dimensions, which capture the most significant features of the data. By limiting the number of latent dimensions involved in the exploration, we make the process more manageable and intuitive for the user.

In particular, we prioritize latent dimensions that exhibit high variance as they tend to encode more meaningful and diverse variations in the data. Dimensions with low variance often contain less useful information and may only represent noise or insignificant fluctuations. By focusing on the most informative dimensions, we allow users to explore the most relevant and semantically rich parts of the latent space, thus enhancing interpretability and reducing unnecessary complexity.

We use a ranking method based on the Structural Similarity Index Measure (SSIM) to achieve this. Specifically, we generate two images for each latent dimension by decoding latent vectors with the dimension set to its extreme values: -3 and 3. Such value limits are selected because they encompass most of the latent space's meaningful variations. As the latent dimensions in β-VAE models are typically assumed to follow a Gaussian distribution, setting the boundaries to -3 and 3 ensures we focus on the latent space's most relevant and stable parts. This approach helps avoid exploring extreme values that may result in unrealistic or irrelevant image reconstructions, as these regions often correspond to noise or outliers.

Once the two extreme images for each latent dimension are generated, we compute the SSIM between them. The SSIM measure evaluates the image's perceptual similarity, which reflects the degree of variation introduced by adjusting a particular latent dimension. A higher SSIM value indicates that the two images are similar, suggesting that the dimension does not introduce significant changes. In contrast, a lower SSIM value suggests that the dimension generates noticeable differences in the images, revealing a dimension that captures meaningful semantic variation in the data.

These SSIM values are then used to rank the latent dimensions. Dimensions that introduce more variation—those with lower SSIM values—are prioritized for user interaction, as they are more likely to provide insightful and semantically rich transformations. By focusing on these ranked dimensions, we ensure that users engage with the most informative and semantically relevant features of the latent space, making the exploration both effective and efficient.

4.4. Dimensionality Reduction and Trajectory Analysis

To better analyze and interpret the high-dimensional latent space, we applied Uniform Manifold Approximation and Projection (UMAP), a popular dimensionality reduction technique. UMAP is particularly effective in preserving the local structure of the data, making it an ideal tool for visualizing and understanding complex relationships in the latent space. By projecting the latent representations into a 2D space, UMAP simplifies the analysis of the high-dimensional data, revealing patterns and clusters that would otherwise be difficult to identify.

One of the key advantages of using UMAP for dimensionality reduction is its ability to uncover natural clusters within the latent space. In datasets such as MNIST, Fashion MNIST, or MedMNIST, UMAP often reveals clusters that correspond to distinct categories, such as different digits or types of clothing. These clusters provide important insights into how the model organizes and encodes the data. For instance, in the case of image datasets, we can visually observe how different categories are grouped together in the latent space, highlighting areas where the model distinguishes between different classes. This is valuable for understanding how the model perceives and organizes the features of the data.

UMAP also plays an important role in visualizing user interactions in the Identikit Game. As users explore the latent space, their movements—adjusting the latent dimensions—can be traced and represented as trajectories in the 2D space. These trajectories reflect the paths users take to move from the starting image to the target image, providing a visual representation of how users are navigating the latent space. The trajectory analysis reveals interesting patterns, such as frequent revisits to specific regions or the types of areas in the latent space that users find more challenging to explore. This insight is critical for improving the user interface and interaction design, as it helps pinpoint where users might need additional guidance or where the latent space could be better structured for easier navigation.

Additionally, by measuring the distance traveled in the latent space during each step of the game, we can assess how efficiently users are exploring the latent dimensions. The distance metric quantifies how much the latent vector changes as users adjust it, providing a measure of the movement required to get closer to the target image. This metric is useful for determining when the user has reached a satisfactory transformation and when the game can be considered completed. By setting a threshold on the distance traveled, we can stop the exploration once the user has made sufficient progress toward the target image, improving the overall efficiency of the interaction.

The combination of UMAP-based dimensionality reduction and trajectory analysis allows for a more intuitive understanding of the latent space. UMAP helps reduce the complexity of the data, making it visually accessible, while trajectory analysis tracks how users move through this space, shedding light on their interactions and decision-making processes. This fusion of techniques enhances the interpretability of the latent space, allowing users to gain deeper insights into how the model encodes and decodes the data.

In high-stakes domains like healthcare, this level of interpretability is essential. The ability to visualize and understand how latent dimensions relate to different categories or outcomes can help clinicians and other experts make more informed decisions. The use of UMAP and trajectory analysis thus not only facilitates the exploration of latent spaces but also ensures that this exploration is meaningful, guiding users toward relevant dimensions and interpretations.

4.5. Reverse Engineering Optimal Latent Dimensions

Another contribution of this work is the proposal of a human-centered metric to determine the optimal number of latent dimensions. By evaluating user performance in the Identikit Game across different latent dimension sizes, we can identify the configuration that maximizes usability and interpretability. Specifically, the optimal latent space minimizes correlations among dimensions while preserving semantic richness while, user performance averaged across domain experts, serves as the primary metric. Higher performance indicates that the latent space is sufficiently expressive without being overly complex.

Preliminary Observations. While human user studies are planned for future work, our experiments yielded several preliminary insights:

- **Latent Dimension Selection:** The SSIM-based ranking effectively identified semantically meaningful dimensions, as confirmed by qualitative analysis of the decoded images.
- **Clustering in UMAP Space:** The 2D projections consistently revealed distinct clusters corresponding to different classes.
- **Trajectory Analysis:** The trajectories in the 2D space provided intuitive visualizations of user interaction, allowing us to identify patterns such as frequent revisits to specific clusters or regions.

Experiments with human users, including medical professionals for the **MedMNIST** dataset, are planned as part of a follow-up study. These studies will provide critical insights into the usability and effectiveness of the interface in real-world scenarios. Additionally, we aim to refine the trajectory-based metrics and integrate them into the game mechanics for automated feedback during interaction.

5. Conclusion and Future Work

In this work, we proposed an interactive visual interface for latent space exploration, integrating human-guided navigation and visualization tools to improve explainability in black-box AI models. Our results demonstrate the potential of the tool to make latent spaces more accessible and interpretable, particularly through the Identikit Game and SSIM-based latent dimension ranking. By allowing users to intuitively manipulate latent dimensions and observe their semantic impact, the interface bridges the gap between technical latent space representations and human understanding.

The implications for explainable AI are significant, especially in medical contexts where trust and accountability are paramount. The application of counterfactual reasoning within this interface offers a powerful tool for generating actionable insights, enabling users to explore ”what-if” scenarios and understand the decision pathways of models. Furthermore, integrating dimensionality reduction techniques, such as UMAP, provides a complementary layer of interpretability by visualizing user trajectories and latent space structure in a 2D plane.

While our interface is designed for intuitive interaction, its effectiveness has not yet been validated through extensive user studies, particularly with domain experts in critical fields such as medicine.

Looking ahead, there are several promising directions for future research: Conducting User studies with medical professionals and other domain experts to evaluate the interface's usability and impact on decision-making processes. We are also planning to improve the user interface to provide real-time feedback, personalized interaction, and adaptive guidance based on user performance. Another goal concerns the expansion of the interface to support a broader range of datasets, including multi-modal and time-series data and to test its generalizability and robustness. More sophisticated metrics need to be developed to quantify user interaction and latent space quality, enabling data-driven optimization of the tool. Eventually, exploring the integration of AI-driven suggestions to guide users through latent space navigation will improve efficiency and reduce cognitive load.

In conclusion, our work represents a step forward in human-centred AI, offering a novel approach to exploring and interpreting latent spaces. By empowering users to interact with abstract, semantically rich representations, we pave the way for more transparent, explainable, and trustworthy AI systems, particularly in high-stakes domains like healthcare. We imagine that this visual interface will serve as a foundation for future innovations in interactive explainability and human-AI collaboration.

6. Acknowledgments

This work has been supported by the Partnership Extended PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”, by TANGO project, grant agreement no.101120763, and by SoBigData.it that receives funding from European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: “SoBigData.it –

Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[2] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature medicine 25 (2019) 44–56.
[3] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).
[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.
[5] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.
[6] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, beta-vae: Learning basic visual concepts with a constrained variational framework, arXiv preprint arXiv:1611.02731 (2017).
[7] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.
[8] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[9] A. Holzinger, G. Langs, D. Denk, K. Zatloukal, H. Müller, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923 (2017).
[10] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 30 (2017).
[11] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harvard journal of law & technology 31 (2017) 841–887.
[12] B. Kim, R. Khanna, O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, Advances in Neural Information Processing Systems 29 (2016) 2280–2288.
[13] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling, in: 2021 IEEE Symposium on Computers and Communications (ISCC), 2021, pp. 1–7. doi:10.1109/ISCC53001.2021.9631485.
[14] C. Metta, A. Beretta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, F. Giannotti, Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning, International Journal of Data Science and Analytics (2023).
[15] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, Guidelines for human-ai interaction, Proceedings of the 2019 CHI conference on human factors in computing systems (2019) 1–13.
[16] J. Jenkins, A. Ojo, S. Coupland, Human-centered artificial intelligence: Engaging human rights, social justice, and public interest in the design and use of ai, Journal of Responsible Technology 4 (2021) 100022.
[17] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).
[18] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605.
[19] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, Journal of Open Source Software 3 (2018) 861.
[20] S. Verma, J. P. Dickerson, K. Hines, Counterfactual explanations for machine learning: A review, arXiv preprint arXiv:2010.10596 (2020).
[21] A. Bhattacharya, J. Ooge, G. Stiglic, K. Verbert, Directive explanations for monitoring the risk of diabetes onset: Introducing directive data-centric explanations and combinations to support what-if explorations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI 2023, ACM, 2023, pp. 204–219. doi:10.1145/3581641.3584075.

[22] M. Riveiro, S. Thill, "that's (not) the output I expected!" on the role of end user expectations in creating explanations of AI systems, Artif. Intell. 298 (2021) 103507. URL: https://doi.org/10.1016/j.artint.2021.103507. doi:10.1016/J.ARTINT.2021.103507.

[23] R. Guidotti, A. Monreale, S. Matwin, D. Pedreschi, Black box explanation by learning image exemplars in the latent feature space, in: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019, Springer, 2019, pp. 189–205.

[24] C. Li, J. Thijssen, T. Abdelaal, J. Tanevski, V. van Unen, B. P. Lelieveldt, T. Hollt, Spacewalker: Interactive gradient exploration for spatial transcriptomics data, bioRxiv (2023).

[25] B. C. Kwon, S. Friedman, K. Xu, S. A. Lubitz, A. Philippakis, P. Batra, P. T. Ellinor, K. Ng, Latent space explorer: Visual analytics for multimodal latent space exploration, arXiv preprint arXiv:2312.00857 (2023).

[26] A. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: from counterfactual explanations to interventions, in: M. C. Elish, W. Isaac, R. S. Zemel (Eds.), FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, ACM, 2021, pp. 353–362. URL: https://doi.org/10.1145/3442188.3445899. doi:10.1145/3442188.3445899.

[27] J. Ooge, K. Verbert, Explaining artificial intelligence with tailored interactive visualisations, in: Companion Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22 Companion, Association for Computing Machinery, New York, NY, USA, 2022, p. 120–123. URL: https://doi.org/10.1145/3490100.3516481. doi:10.1145/3490100.3516481.

[28] D. Wang, Q. Yang, A. M. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable AI, in: S. A. Brewster, G. Fitzpatrick, A. L. Cox, V. Kostakos (Eds.), Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019, ACM, 2019, p. 601. URL: https://doi.org/10.1145/3290605.3300831. doi:10.1145/3290605.3300831.

[29] B. Montambault, G. Appleby, J. Rogers, C. D. Brumar, M. Li, R. Chang, Dimbridge: Interactive explanation of visual patterns in dimensionality reductions with predicate logic, IEEE Trans. Vis. Comput. Graph. 31 (2025) 207–217. URL: https://doi.org/10.1109/TVCG.2024.3456391. doi:10.1109/TVCG.2024.3456391.

[30] Q. V. Liao, K. R. Varshney, Human-centered explainable AI (XAI): from algorithms to user experiences, CoRR abs/2110.10790 (2021). URL: https://arxiv.org/abs/2110.10790. arXiv:2110.10790.

[31] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38. URL: https://doi.org/10.1016/j.artint.2018.07.007. doi:10.1016/J.ARTINT.2018.07.007.

[32] M. Guesmi, M. A. Chatti, S. A. Joarder, Q. U. Ain, R. Alatrash, C. Siepmann, T. Vahidi, Interactive explanation with varying level of details in an explainable scientific literature recommender system, Int. J. Hum. Comput. Interact. 40 (2024) 7248–7269. URL: https://doi.org/10.1080/10447318.2023.2262797. doi:10.1080/10447318.2023.2262797.

[33] A. M. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. S. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: R. L. Mandryk, M. Hancock, M. Perry, A. L. Cox (Eds.), Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018, ACM, 2018, p. 582. URL: https://doi.org/10.1145/3173574.3174156. doi:10.1145/3173574.3174156.