P1 — Hierarchical Concept-Based Explainable-by-Design Models

Explainable and Trustworthy AI Course

Politecnico di Torino - 2025/2026

Reference teachers: Eliana Pastor, Eleonora Poeta

Project. The project aims to design and evaluate concept-based explainable-by-design models in which concepts are explicitly organized in a hierarchy, enabling explanations and predictions at multiple levels of abstraction.

Overview.

Concept-based explainable models aim to improve interpretability by expressing predictions in terms of human-understandable concepts [4]. In most existing approaches, however, concepts are represented as a flat set of human-defined symbols. This simplification overlooks that concepts are naturally organized in hierarchies, from general to more specific ones. For example, specific concepts such as *beak* can be organized into higher-level parts such as *head*.

Hierarchical structures over concepts can be obtained in different ways. For symbolic concepts, hierarchies can be derived from existing ontologies, such as WordNet, which define relationships between concepts and organize specific concepts into more general ones. Alternatively, hierarchical structures can be automatically constructed, for instance by querying a large language model, extending recent approaches that generate flat concept sets by introducing relations among concepts.

Recent works have started to explore hierarchical structures in concept-based models and interpretable architectures, showing that multi-level representations can improve interpretability and predictive performance [2, 3, 5]. Moreover, hierarchical reasoning enables explanations at different levels of granularity and supports a more nuanced understanding of model behavior. For example, a model may correctly identify a high-level category even when it fails at a fine-grained classification. In parallel, recent research has highlighted that not all classification errors are equally severe: confusing semantically similar classes is less problematic than confusing unrelated ones [1]. Hierarchical relationships between classes can be used to quantify the severity of mistakes and design more informative evaluation strategies. Despite these advances, current concept-based explainable models rarely integrate hierarchical structures explicitly, and there is still a lack of unified frameworks that combine hierarchical concept representations, interpretable predictions, and hierarchy-aware evaluation.

Goal.

The goal of this project is to investigate hierarchical concept-based explainable models by designing a framework in which concepts are structured across multiple levels of abstraction. The project aims to assess whether hierarchical representations improve interpretability, enable richer explanations, and allow for a more informative analysis of model predictions and errors.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of concept-based explainable models, with particular focus on concept bottleneck models, coarse-to-fine concept discovery, and hierarchical prototype-based methods. Review also hierarchy-aware evaluation methods for classification errors and semantic mistake severity.
- **Identification of Research Gaps.** Identify limitations of current concept-based approaches, such as flat concept representations, insufficient modeling of coarse-to-fine semantic relations, and limited evaluation of mistakes at different abstraction levels.
- **Implementation.** Design and implement a hierarchical concept-based model. Possible directions include:
  - defining a multi-level concept bottleneck architecture, e.g., by extending an existing CBM with *is-a* concept relations;
  - constraining the prediction process so that coarse concepts support fine-grained concepts;
  - generating explanations at multiple levels of granularity.
- **Evaluation.** Evaluate the proposed model on at least one dataset using both predictive and interpretability metrics. The evaluation should include standard predictive performance, concept quality, and hierarchy-aware analysis, for example, by assessing whether the model makes semantically better mistakes or whether explanations become more meaningful across different levels of abstraction.

References

[1] Luca Bertinetto et al. "Making better mistakes: Leveraging class hierarchies with deep networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 12506–12515.

[2] Peter Hase et al. "Interpretable image recognition with hierarchical prototypes". In: *Proceedings of the AAAI conference on human computation and crowdsourcing*. Vol. 7. 2019, pp. 32–40.

[3] Federico Pittino, Vesna Dimitrievska, and Rudolf Heer. "Hierarchical concept bottleneck models for vision and their application to explainable fine classification and tracking". In: *Engineering Applications of Artificial Intelligence* 118 (2023), p. 105674.

[4] Eleonora Poeta et al. "Concept-based Explainable Artificial Intelligence: A Survey". In: *ACM Comput. Surv.* (Nov. 2025). ISSN: 0360-0300. DOI: 10.1145/3774643. URL: https://doi.org/10.1145/3774643.

[5] Ao Sun et al. "Boosting Concept Bottleneck Models with Supervised, Hierarchical Concept Learning". In: *Paper under review.* *https://openreview.net/forum?id=Q9ZOc1Rb5i* ().

P2 — Explainable-by-Design Models for Speech Analysis

Explainable and Trustworthy AI Course

Politecnico di Torino — 2025/2026

Reference teachers: Gabriele Ciravegna, Eleonora Poeta

Project. The project aims to design and evaluate explainable models for speech paralinguistic analysis, enabling interpretable predictions of attributes such as emotion, sentiment, speaker age, or stress level through human-understandable explanations.

Overview.

Self-supervised speech models such as HuBERT [3] and wav2vec 2.0 [1] have achieved remarkable performance across a wide range of paralinguistic tasks, including emotion recognition, sentiment analysis, and speaker attribute estimation. Despite their effectiveness, these models remain largely opaque: they produce predictions without offering any human-interpretable justification. This lack of transparency is especially problematic in high-stakes domains such as healthcare, affective computing, and human-computer interaction, where understanding the rationale behind a prediction is as important as the prediction itself.

Explainability methods for speech models can be broadly categorised into post-hoc approaches—which produce explanations for an already-trained black-box model—and explainable-by-design approaches—which embed interpretability directly into the model architecture [5]. On the post-hoc side, attribution methods such as SHAP or gradient-based saliency can identify which parts of the input signal or which features most influence a prediction, while counterfactual explanations describe the minimal change to the input that would alter the output. On the by-design side, Concept Bottleneck Models [4] constrain predictions to pass through an intermediate layer of human-interpretable concepts (e.g., *high pitch*, *fast speech rate*, *breathy voice quality*), enabling practitioners to inspect concept activations and perform targeted interventions [7]. Attention-based models offer an alternative perspective by providing a breakdown of the most salient frames or tokens influencing the final decision.

A common challenge across all these paradigms is grounding explanations in the acoustic and paralinguistic properties of speech. Unlike vision, where spatial saliency maps are intuitive, or text, where token importance is straightforward, speech explanations must be meaningful in terms of the signal's temporal and spectral structure. Concept sources for by-design models include low-level acoustic descriptors (fundamental frequency, energy, MFCCs), higher-level

prosodic attributes (speaking rate, pauses, rhythm), or semantic descriptions automatically generated by large language models. Despite recent advances, a comprehensive study of explainability methods across the broader paralinguistic landscape—covering multiple target attributes, diverse explanation types, and user-facing evaluation—remains largely missing.

Goal.

The goal of this project is to design and evaluate explainable models for speech paralinguistic analysis. Students will investigate how different explainability paradigms (post-hoc or by-design) can be applied to pre-trained speech representations, and assess the trade-offs between predictive performance and interpretability. The project aims to demonstrate the practical value of explanations in a speech analysis scenario, providing insights into the features and patterns that drive paralinguistic predictions.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of explainability methods for deep learning (post-hoc and by-design), pre-trained self-supervised speech models, and paralinguistic analysis tasks (emotion, sentiment, speaker attributes). Include a survey of existing XAI methods applied to speech and audio.
- **Identification of Research Gaps.** Identify limitations of current black-box speech models and the specific challenges of producing meaningful explanations over acoustic signals, such as the difficulty of defining intuitive input features, the absence of standard concept taxonomies for paralinguistics, and the lack of user-facing evaluation of speech explanations.
- **Implementation.** Design and implement at least one explainable model for a paralinguistic task, using a pre-trained speech representation (e.g., HuBERT or wav2vec 2.0) as a backbone. Possible directions include:
  - a post-hoc attribution method identifying the most influential acoustic features or temporal segments for a given prediction;
  - a Concept Bottleneck Model predicting a set of interpretable speech concepts before the final classifier, enabling concept inspection and intervention;
  - an attention-based model providing a temporal breakdown of the salient frames driving the prediction.

  The model should be applied to at least one benchmark dataset, such as IEMOCAP [2] or MELD [6].
- **Evaluation.** Compare the explainable model against a black-box baseline in terms of predictive performance. Evaluate the quality of explanations using faithfulness metrics, alignment with domain knowledge, or

human assessment. Where applicable, demonstrate the effect of targeted interventions to illustrate the practical value of the chosen explainability approach.

References

[1] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 12449–12460.

[2] Carlos Busso et al. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* 42.4 (2008), pp. 335–359.

[3] Wei-Ning Hsu et al. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3451–3460.

[4] Pang Wei Koh et al. "Concept bottleneck models". In: *International conference on machine learning*. PMLR. 2020, pp. 5338–5348.

[5] Eleonora Poeta et al. "Concept-based Explainable Artificial Intelligence: A Survey". In: *ACM Comput. Surv.* (Nov. 2025). ISSN: 0360-0300. DOI: 10.1145/3774643. URL: https://doi.org/10.1145/3774643.

[6] Soujanya Poria et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 527–536.

[7] Mateo Espinosa Zarlenga et al. "Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.

P3 — Mechanistic Interpretability of Vision Transformers

Explainable and Trustworthy AI Course

Politecnico di Torino — 2025/2026

Reference teachers: Gabriele Ciravegna, Eliana Pastor

Project. The project aims to investigate the internal representations of Vision Transformers by applying mechanistic interpretability techniques—such as Sparse Autoencoder feature decomposition or circuit-level causal analysis—in order to understand what visual features and computational structures drive model predictions.

Overview.

Vision Transformers (ViTs) [4] have become the dominant architecture for visual recognition, achieving state-of-the-art performance on image classification and beyond. Despite their empirical success, the internal computations of ViTs remain poorly understood. Identifying which features a model relies on, and how information flows from input patches to the final prediction, is a key open problem in interpretable deep learning.

Mechanistic interpretability is a research program that aims to reverse-engineer neural networks by identifying the computational structures responsible for specific model behaviors [5]. Two complementary tools have proven particularly productive in the context of transformer-based language models. First, *circuit analysis* identifies the specific attention heads and MLP layers that together implement a given capability, by systematically measuring the causal contribution of each component through activation patching [2]. Second, *Sparse Autoencoders* (SAEs) decompose the high-dimensional, polysemantic activation space of a model into a larger set of sparse, interpretable directions, each of which tends to correspond to a recognizable, monosemantic concept [1, 3].

While mechanistic interpretability has yielded striking findings for large language models—including the discovery of circuits for indirect object identification, factual recall, and in-context learning—its application to Vision Transformers is still nascent. ViTs present unique challenges compared to language models: their tokens represent spatial image patches rather than words, information aggregates into a class token via attention over space, and polysemanticity in vision activations may take different forms. Preliminary work has begun to characterise the roles of attention heads in ViTs [6], but a systematic mechanistic study of vision models remains an open direction.

Goal.

The goal of this project is to apply mechanistic interpretability techniques to a pre-trained Vision Transformer, with the aim of understanding what visual features and computational structures emerge in its internal representations. Students will select at least one of the two main mechanistic interpretability paradigms—SAE-based feature decomposition or circuit-level causal analysis—and contribute to a principled characterisation of how ViTs process visual information.

Required analysis, implementation, and evaluation.

- **Literature Review.** Review the Vision Transformer architecture and pre-training paradigms (supervised ViT, DINO). Study mechanistic interpretability for transformers, covering transformer circuits [5], SAEs [1, 3], and automated circuit discovery [2]. Survey existing work on interpreting ViT internals [6].
- **Identification of Research Gaps.** Identify how the spatial structure of ViTs (patch tokens, class token aggregation, bidirectional attention) creates opportunities and challenges relative to language models. Highlight the absence of systematic SAE and circuit analyses for vision models as a key gap in the mechanistic interpretability literature.
- **Implementation.** Working from a pre-trained ViT backbone (e.g., ViT-B/16 on ImageNet or a DINO ViT), implement one of the following directions:
  - *Sparse Autoencoders*: train one or more SAEs on the residual stream or MLP output activations of selected ViT layers, and identify the visual concepts encoded by the learned sparse features;
  - *Circuit analysis*: use activation patching to isolate the attention heads and MLP components causally responsible for a specific classification behavior on a controlled task or image category.
- **Evaluation.** Assess the interpretability of SAE-learned features (if applicable) through automated labeling via CLIP or an LLM, and through human evaluation. Measure circuit faithfulness (if applicable) by ablating identified components and observing the resulting drop in task performance. Analyse how features or circuits vary across layers to characterise the progression of visual representations through the network.

References

[1] Trenton Bricken et al. "Towards monosemanticity: Decomposing language models with dictionary learning". In: *Transformer Circuits Thread* 2.5 (2023), p. 6.

[2] Arthur Conmy et al. "Towards Automated Circuit Discovery for Mechanistic Interpretability". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.

[3] Hoagy Cunningham et al. "Sparse Autoencoders Find Highly Interpretable Features in Language Models". In: *International Conference on Learning Representations*. 2024.

[4] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[5] Nelson Elhage et al. "A Mathematical Framework for Transformer Circuits". In: *Transformer Circuits Thread* (2021). URL: `https://transformer-circuits.pub/2021/framework/index.html`.

[6] Maithra Raghu et al. "Do Vision Transformers See Like Convolutional Neural Networks?" In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

P4 — Evaluating the Quality of Explanations

Explainable and Trustworthy AI Course

Politecnico di Torino — 2025/2026

Reference teachers: Eliana Pastor, Eleonora Poeta

Project. This project aims to explore the evaluation of explanation methods, focusing on how to assess the quality, reliability, and usefulness of explanations provided by explainable AI techniques. In addition to analyzing existing approaches, the project will involve proposing new evaluation strategies to better capture different aspects of explanation quality.

Overview.

Explanation methods are designed to provide insights into machine learning model decisions. However, evaluating the quality of these explanations remains a challenging and open problem. While many approaches exist [2, 4, 5, 7, 6, 9, 13, 12, 14], they often focus on specific aspects such as robustness or sensitivity, without providing a comprehensive view of explanation quality. Explanations should be assessed across multiple dimensions, including faithfulness to the model, plausibility for humans, robustness to perturbations, and usability in real-world contexts. A systematic evaluation is therefore needed to move from anecdotal evidence toward quantitative and comparable assessment of explanations.

Different evaluation criteria capture complementary aspects of explanation quality. For instance, faithfulness assesses whether the explanation reflects the true behavior of the model, while plausibility evaluates whether it aligns with human reasoning. Other properties, such as consistency, continuity, and compactness, further contribute to understanding the reliability and usefulness of explanations. Despite the variety of proposed metrics and evaluation frameworks [10, 3, 1, 8, 11], existing tools often cover only a subset of these aspects, and there is still no unified approach to evaluate explanations in a comprehensive and systematic way. This project aims not only to analyze these limitations but also to explore novel evaluation strategies that better capture the multi-faceted nature of explanation quality.

Goal.

The goal of this project is to study and systematize the evaluation of explanation methods by analyzing multiple dimensions of explanation quality. The project aims to identify limitations of current evaluation strategies and to propose new or improved evaluation methods that provide a more comprehensive and meaningful assessment of explanations.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing evaluation methods for explanation quality. The review should cover different dimensions such as faithfulness, plausibility, robustness, stability, and compactness, as well as different evaluation paradigms, including perturbation-based methods, sanity checks, synthetic data evaluations, and user studies. The review should analyze how these methods capture different aspects of explanation quality and where they fall short.

- **Identification of Research Gaps.** Identify key limitations in current evaluation approaches. For example, many methods focus on a single property of explanations, such as robustness or sensitivity, without considering other aspects such as usability or human interpretability. Other challenges include the lack of standardized benchmarks, the difficulty of comparing different explanation methods, and the gap between quantitative metrics and human-centered evaluation. Based on this analysis, define a clear direction for improving explanation evaluation

- **Implementation.** Select a set of explanation methods and evaluation techniques to analyze. Design and implement an experimental framework that allows evaluating explanations across multiple dimensions. As part of this phase, propose and implement at least one novel evaluation method or metric, or extend an existing one to better capture an aspect of explanation quality that is currently underexplored (e.g., combining faithfulness and plausibility, or integrating user-oriented criteria).

- **Evaluation.** Assess the behavior of explanation methods under the proposed evaluation framework. Compare existing metrics with the proposed evaluation approach. Evaluate whether the proposed evaluation framework and metric(s) provide additional insights compared to existing metrics and discuss their strengths, limitations, and applicability.

References

[1] Chirag Agarwal et al. "Openxai: Towards a transparent evaluation of model explanations". In: *Advances in neural information processing systems* 35 (2022), pp. 15784–15799.

[2] Elvio Amparore, Alan Perotti, and Paolo Bajardi. "To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods". In: *PeerJ Computer Science* 7 (2021), e479.

[3] Giuseppe Attanasio et al. "ferret: a Framework for Benchmarking Explainers on Transformers". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 2023, pp. 256–266.

[4] Francesco Bodria et al. “Benchmarking and survey of explanation methods for black box models”. In: *Data Mining and Knowledge Discovery* 37.5 (2023), pp. 1719–1778.

[5] Oana-Maria Camburu et al. “Can I trust the explainer? Verifying post-hoc explanatory methods”. In: *arXiv preprint arXiv:1910.02065* (2019).

[6] Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. “Framework for evaluating faithfulness of local explanations”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 4794–4815.

[7] Ann-Kathrin Dombrowski et al. “Towards robust explanations for deep neural networks”. In: *Pattern Recognition* 121 (2022), p. 108194.

[8] Anna Hedström et al. “Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond”. In: *Journal of Machine Learning Research* 24.34 (2023), pp. 1–11.

[9] Cheng-Yu Hsieh et al. “Evaluations and methods for explanation through robustness analysis”. In: *arXiv preprint arXiv:2006.00442* (2020).

[10] Phuong Quynh Le et al. “Benchmarking eXplainable AI-A Survey on Available Toolkits and Open Challenges.” In: *IJCAI*. 2023, pp. 6665–6673.

[11] Yang Liu et al. “Synthetic Benchmarks for Scientific Research in Explainable Machine Learning”. In: *Advances in Neural Information Processing Systems Datasets Track*. 2021.

[12] Giambattista Parascandolo et al. “Learning explanations that are hard to vary”. In: *arXiv preprint arXiv:2009.00329* (2020).

[13] Dylan Slack et al. “Reliable post hoc explanations: Modeling uncertainty in explainability”. In: *Advances in neural information processing systems* 34 (2021), pp. 9391–9404.

[14] Chih-Kuan Yeh et al. “On the (in) fidelity and sensitivity of explanations”. In: *Advances in neural information processing systems* 32 (2019).

P5 — Unsupervised Concept Discovery and Evaluation for Medical Vision–Language Models

Explainable and Trustworthy AI Course

Politecnico di Torino - 2025/2026

Reference teachers: Eleonora Poeta, Eliana Pastor

Project. This project investigates unsupervised concept discovery and quantitative evaluation of interpretability in medical Vision–Language Models, building upon the MedConcept framework.

Overview.

Medical Vision–Language Models (VLMs) have recently achieved strong performance across a variety of clinical tasks, including disease classification, segmentation, and report generation. However, these models rely on high-dimensional and polysemantic latent representations that remain largely opaque, limiting their trustworthiness in safety-critical applications [4]. Traditional interpretability techniques, such as gradient-based saliency maps or attention visualizations, provide localized explanations but often fail to capture the semantic structure of learned representations and are typically tied to specific tasks [7]. As a result, they offer limited insight into how models internally organize medical knowledge.

Concept-based explainability has emerged as a promising direction to address these limitations by expressing predictions in terms of human-understandable concepts [8, 9]. More recently, unsupervised approaches have been proposed to automatically discover such concepts from pretrained representations without relying on manual annotations. In particular, sparse autoencoders have shown strong potential for disentangling neural representations into interpretable and semantically meaningful features [3, 11].

Building on these ideas, the MedConcept framework introduces an unsupervised pipeline for extracting latent medical concepts from pretrained VLMs and grounding them in clinically meaningful textual semantics [6]. The framework leverages sparse autoencoders to identify neuron-level concept activations and aligns them with a curated medical vocabulary derived from ontologies such as UMLS [2]. Importantly, concept discovery is performed without supervision, making it applicable across tasks and datasets.

A major challenge in concept-based interpretability is evaluation. Prior work often relies on qualitative inspection, which is subjective and difficult to scale. To address this issue, recent approaches propose using large language models (LLMs) as external evaluators to assess semantic alignment between predicted concepts and textual evidence. In MedConcept, this idea is operationalized

through three quantitative metrics—*Aligned*, *Unaligned*, and *Uncertain*—which capture semantic agreement, contradiction, and ambiguity with respect to clinical reports [6].

Despite these advances, several open challenges remain. Automatically discovered concepts may lack robustness or clinical validity, evaluation may be biased by incomplete or selective clinical reports, and the reliance on pretrained VLMs and LLMs introduces potential sources of bias and misalignment. Moreover, current approaches typically treat concepts as independent entities, overlooking the structured relationships that naturally exist among medical concepts, such as anatomical hierarchies or causal dependencies.

Goal.

The goal of this project is to study unsupervised concept discovery in medical VLMs and critically analyze how such concepts can be used to improve interpretability. The project will explore both the extraction of latent concepts and the design of reliable evaluation strategies, with particular focus on the role of external language models in assessing semantic alignment.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of interpretability methods for medical Vision–Language Models, with particular focus on concept-based explainability approaches such as Concept Bottleneck Models and TCAV [8], as well as recent advances in unsupervised concept discovery using sparse autoencoders and representation disentanglement [12, 5, 1, 10]. The review should also cover existing evaluation strategies for explanations, including LLM-based semantic evaluation protocols and their limitations [13].

- **Identification of Research Gaps.** Identify key limitations of current approaches, such as the lack of guarantees on the semantic validity and stability of discovered concepts, the dependence on external resources such as ontologies and large language models, and the reliance on incomplete or noisy textual reports for evaluation. Particular attention should be given to the absence of structured relationships among concepts and the challenges in quantitatively assessing interpretability in a scalable and reproducible way.

- **Implementation.** Design and implement a simplified unsupervised concept discovery pipeline inspired by recent frameworks. Possible directions include:

  - training a sparse autoencoder to extract latent concepts from pretrained visual or multimodal embeddings;
  - aligning latent features with textual concepts using similarity-based matching within a shared embedding space;

  - generating concept-based explanations for individual samples by identifying highly activated concepts;
  - exploring simple extensions such as filtering, clustering, or organizing discovered concepts into structured representations.

- **Evaluation.** Evaluate the proposed approach on at least one dataset using both qualitative and quantitative interpretability metrics. The evaluation should include analysis of concept quality and semantic alignment, for example by implementing or adapting metrics such as Aligned, Unaligned, and Uncertain scores, and by assessing the consistency and usefulness of the generated explanations. Additional analysis may investigate sensitivity to model design choices and characterize failure cases such as ambiguous or spurious concepts.

References

[1] Usha Bhalla et al. "Interpreting clip with sparse linear concept embeddings (splice)". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 84298–84328.

[2] Olivier Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D267–D270.

[3] Trenton Bricken et al. "Towards monosemanticity: Decomposing language models with dictionary learning". In: *Transformer Circuits Thread* 2.5 (2023), p. 6.

[4] Tribikram Dhar et al. "Challenges of deep learning in medical image analysis—improving explainability and trust". In: *IEEE Transactions on Technology and Society* 4.1 (2023), pp. 68–75.

[5] Shizhan Gong et al. "Concepts from Neurons: Building Interpretable Medical Image Diagnostic Models by Dissecting Opaque Neural Networks". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2025, pp. 3–18.

[6] Md Rakibul Haque et al. "MedConcept: Unsupervised Concept Discovery for Interpretability in Medical VLMs". In: *arXiv preprint arXiv:2604.11868* (2026).

[7] Daniel T Huff, Amy J Weisman, and Robert Jeraj. "Interpretation and visualization techniques for deep learning models in medical imaging". In: *Physics in Medicine & Biology* 66.4 (2021), 04TR01.

[8] Been Kim et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)". In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.

[9] Pang Wei Koh et al. "Concept bottleneck models". In: *International conference on machine learning*. PMLR. 2020, pp. 5338–5348.

[10] Chris Olah et al. “Zoom in: An introduction to circuits”. In: *Distill* 5.3 (2020), e00024–001.

[11] Mateusz Pach et al. “Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models”. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2026. URL: https://openreview.net/forum?id=DaNnkQJSQf.

[12] Sukrut Rao et al. “Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 444–461.

[13] Lin Shi et al. “Judging the judges: A systematic study of position bias in llm-as-a-judge”. In: *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. 2025, pp. 292–314.