SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Concept Bottleneck Optimization Algorithms For High Fidelity Medical Image Interpretation

Dr.U. Premkumar[1*], Dr.B.H. Parameshwar Keerthi[2], Dr. Parvindar Kaur Chhabda[3], Dr. Chikati Madhava Rao[4], Abdullayeva Shakhnoza Anvarovna[5]

[1*]Professor, Department of Radio Diagnosis, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Enathur, Kanchipuram, Tamil Nadu, India. E-mail: premku@maher.ac.in,

[2]Assistant Professor, Department of Radio Diagnosis, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Enathur, Kanchipuram, Tamil Nadu, India. E-mail: keerthiparam@maher.ac.in

[3]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.parvindarkaurchhabda@kalingauniversity.ac.in, https://orcid.org/0009-0006-9775-7254

[4]Associate Professor, Department of CSE, CMR Technical Campus, Hyderabad, Telangana, India. E-mail: madhav.chikati@gmail.com

[5]Turan International University, Namangan, Uzbekistan. E-mail: shaxnoza.abdullayeva.80@mail.ru, https://orcid.org/0009-0004-4826-0175

*Corresponding author: Email: premku@maher.ac.in

Abstract

Interpretable prediction in high-stakes medical imaging tasks is a promising paradigm that can be achieved through Concept Bottleneck Models (CBMs). While promising in theory, the current CBMs suffer from conceptual lacunae, annotation limitations, and suboptimal convergence in complex multi-pathology scenarios, all of which reduce diagnostic fidelity. The research proposes the Concept Bottleneck Optimization Algorithm (CBOA), a novel framework that combines adaptive multi-scale feature extraction, semantic concept alignment, gradient-based concept pruning, and a residual concept learner that learns concepts to boost concept detection accuracy and downstream diagnostic performance simultaneously. The CBOA achieves state-of-the-art AUC scores of 95.6% for the NIH Chest X-ray14 and 94.9% for the ISIC 2020 dermoscopy corpus, with concept accuracy scores of 91.8% and 90.6%, respectively. Comparative analyses with standard deep learning classifiers and previous CBM variants demonstrate statistical significance for both CBOA's interpretability and classification fidelity. The algorithm addresses the known problem of concept rigidity in predefined concepts by adding a complementary residual pathway that discovers latent concepts not in the manually annotated list. The results confirm the clinical feasibility of the CBOA system for accurate and transparent medical image diagnosis.

Keywords: Concept Bottleneck Models; Medical Image Interpretation; Explainable Artificial Intelligence; Gradient Pruning; Vision Transformer; Diagnostic Fidelity; Semantic Alignment.

1. Introduction

In the medical fields of clinical radiology, dermatology, and pathological diagnosis, the use of deep learning systems has reached the level of performance of an experienced clinician on specific tasks [1]. However, the transparency of typical convolutional and transformer models makes them inapplicable in scenarios where clinicians need to provide an explanation, audit and defend a diagnostic decision [2]. As a result, there is a constant trust challenge between the capabilities of algorithms and the adoption in the clinical scenario, which has led to the active research sub-field of Explainable Artificial Intelligence (XAI) in medical imaging [3]. One of the most prominent approaches to XAI are the Concept Bottleneck Models (CBMs), which insert an explicit and understandable bottleneck layer between the image encoder and the classification head [4]. A CBM first anticipates whether a visual concept (e.g., nodule spiculation in computed tomography or asymmetric pigment networks in dermoscopy) is present and how intense it is, and then infers the diagnostic class based on these predicted concept activations alone. This architecture gives the model two desirable features: (i) a clear reasoning chain that is easily readable by clinicians, and (ii) the ability of human intervention at the concept level without retraining the entire network [5]. However, there are some unanswered questions regarding existing CBMs. First, it is rarely possible

to guarantee that the concept is complete; that is, if the concept is underspecified, the residual diagnostic information will be thrown in the way of the concept scores, which will limit the scope of interpretation [6]. Second, most clinical datasets are too small to afford the time and expense of manual concept-level annotation [7]. Third, concept prediction and task prediction are often confounded by gradient interference, making it hard for the concept layer to discover spurious correlations instead of clinically relevant features [8]. Fourth, the standard CBMs are trained using single-scale representations, while the medical pathologies occur at several spatial scales at the same time [9]. This paper provides solutions to all the four challenges by implementing the Concept Bottleneck Optimization Algorithm (CBOA). The key findings of this study are summarized below: An adaptive multi-scale feature extraction module, based upon a Vision Transformer (ViT-L/16) backbone with both fine-grained and coarse semantic cues. • Semantic concept alignment based on a semantic concept dictionary generated by a large language model (LLM), enabling to go beyond manually annotated label sets and saving annotation efforts without compromising the grounding of concepts. A concept pruning method based on gradient, which removes redundant and clinically irrelevant concepts gradually in the iterative process to eliminate the gradient interference and stabilize training process. • Residual concept learner that learns and encodes latent concepts that are not pre-defined in the dictionary, without demanding exhaustive expert annotation of concepts. The CBOA shows consistently improved performance over well-established baselines across diagnostic and concept fidelity measures, both in the NIH Chest X-ray14 [10] and ISIC 2020 [11] benchmarks, and maintains outputs suitable for clinicians.

This paper is organized as follows: Section 2 presents a review of related work; Section 3 presents the proposed CBOA methodology; Section 4 details the experimental setup; Section 5 presents both quantitative and qualitative results; Section 6 discusses implications and limitations and Section 7 concludes the paper.

2. Related Work

2.1 Concept-Based Interpretability in Medical Imaging

The basic CBM architecture of [4] showed an interpretable, intervention-capable model to predict an intermediate concept vector prior to giving a class label. This paradigm was expanded through subsequent work in a few directions. Post-hoc CBMs [8] inserted concept bottlenecks into pre-trained black-box models, but compromised their concept fidelity in the process of maintaining their predictive performance. Interactive CBMs [12] are CBMs that are augmented with a user correction mechanism, and demonstrate that, for out-of-distribution test cases, clinician intervention at test time significantly increases diagnostic accuracy. [13] proposed visual concept filtering to adaptively suppress concept activation due to imaging artefacts but not pathological tissues, which improved the multi-class chest radiograph classification. Concept complement bottleneck models [6] recently suggested supplementing pre-existed concept set with auto-discovered visual concept, thus creating a connection between manual concept and concept set completion. Multi-label CBM frameworks for chest X-ray classification [14] introduced concept generation without expert labelling per image, using large VLMs. The bottleneck was trained to pay attention to the same features as the attending clinician, thus showing significant improvements in out-of-domain transfer to the white blood cell and skin lesion questions.

2.2 Optimization Challenges in CBMs

Gradient interference continues to be one of the major reasons for degradation in the performance of simultaneously trained CBMs [16]. Sequential training (first concept predictor and then task predictor) helps to reduce interference and sacrifices concept accuracy [4]; Marcinkevicius et al. [17] showed that enforcing orthogonality constraints on the concept representation through concept whitening results in better disentanglement, at the cost of task accuracy. Concept regularisation with adversarial learning [18] has been proposed to stop the task prediction gradient from contaminating the concept representations; there is some success with ultrasound training. There has been relatively little work on the pruning of concepts, which is the removal of concepts that provide no further information or are redundant. [19] reviewed some concept-based model improvement methods and found Gradient-based Saliency methods as most principled approach to concept selection but not the direct integration in an optimisation loop has been systematically explored, before the present research.

2.3 Vision Transformers for Medical Image Analysis

To take advantage of long-range self-attention, Vision Transformers (ViT) have replaced convolutional backbones as the default backbone for high-resolution medical image applications [20]. The concept alignment scores of models pre-trained on large-scale medical corpora with self-supervised learning are significantly higher than those pre-trained on ImageNet, demonstrating that concept-aligned representations are essential for successful concept bottleneck learning [21]. The present work takes advantage of a pre-trained ViT-L/16 model that was trained on the combined CheXpert and ISIC data sets to provide a good baseline of representation in CBOA prior to concept bottleneck fine-tuning.

3. Methodology

The proposed CBOA framework starts with the formalization of the problem, which is to learn a mapping $f: X \rightarrow (p_C, \hat{y})$, where $X \in \mathbb{R}^{H\times W\times C}$, is a medical image, the concept dictionary C is pre-defined with K visual concepts, $C = \{c_1, c_2, \dots, c_K\}$and a diagnostic class label $Y \in \{1, \dots, M\}$, is given. The constraint is that the prediction class $\hat{y}$is based mostly on p_C (modulo the residual concept pathway) and that p_C, is semantically consistent with the dictionary C in (Equation 1).

$$f: X \rightarrow (p_C, \hat{y}), p_C \in [0,1]^K, \hat{y} \in \Delta^M \quad (1)$$

A pathological feature extraction module with adaptive multi-scale feature extraction is used to capture pathological features at multiple spatial scales with a ViT-L/16 image encoder. The attention maps from the last three transformers (layers 22, 23, 24) are projected using light-weight MLPs and combined with learnable weights $w_l \geq$, subject to the constraint of $\sum w_l = 1$, forming the multi-scale feature tensor Z, as shown in equation (2):

$$Z = \sum_{l\in\{22,23,24\}} w_l \cdot \mathrm{MLP}_l(A_l) \quad (2)$$

A prototype embedding $e_k$for each concept $c_k$is computed by averaging the CLS token representations of annotated images. The semantic concept alignment loss promotes similarity when the concept is there, otherwise it promotes dissimilarity:

$$L_{\text{align}} = \sum_k [a_k \cdot \max(0, \delta - \mathrm{sim}(Z, e_k)) + (1 - a_k) \cdot \max(0, \mathrm{sim}(Z, e_k) - \epsilon)] \quad (3)$$

where in equation (3) $a_k \in \{0,1\}$is the ground-truth annotation $\mathrm{sim}(Z, e_k)$denotes cosine similarity, and $\delta, \epsilon$are margin hyperparameters. The importance of concept c_k is then calculated as the mean absolute gradient of the task loss $L_{\text{task}}$with respect to its probability score p_{c_k}:and taken as the value to be pruned, and this is done iteratively:

$$I_k = \mathbb{E}_{(x,y)\in D_{\text{val}}} \left| \frac{\partial L_{\text{task}}}{\partial p_{c_k}} \right| \quad (4)$$

Equation (4) explains the concepts with L_{task}below a dynamic threshold τare masked from gradient propagation in subsequent epochs, reducing gradient interference and focusing learning on diagnostically informative concepts. To compensate for any visual information not contained in the concept dictionary, the residual concept learner Rproduces a vector $p_R \in \mathbb{R}^d$and the final prediction is determined as:

$$\hat{y} = \mathrm{softmax}\left(W_y \cdot [p_C; p_R] + b_y\right) \quad (5)$$

The residual pathway in equation (5) $L_{\text{reg}} = \lambda \| p_R \|_1$, takes into account a regularization term that punishes uninformative residual activations and guarantees that the residual pathway encodes novel and meaningful information. The overall training goal is the combination of task loss, concept prediction loss, alignment loss, and residual regularization:

$$L = L_{\text{task}} + \alpha \cdot L_{\text{concept}} + \beta \cdot L_{\text{align}} + \lambda \cdot L_{\text{reg}} \quad (6)$$

In equation (6) α, β, λare scalar hyperparameters (optimized by grid search), and the model is trained with AdamW optimizer with a cosine annealing learning rate schedule. This design allows CBOA to learn semantically meaningful concept representations together with a compact residual concept space and accurate diagnostic predictions from complex medical images.

4. Experimental Setup

4.1 Datasets

NIH Chest X-ray14 [10]: Chest radiographs with 14 disease classes for 112,120 frontal-view images from 30,805 patients that are publicly available. The official train/validation/test split is followed. The radiological attributes used for concept annotation include 18 attributes that are obtained from the expanded annotation set of Irvin et al. [22] (e.g., pleural effusion sign, cardiomegaly ratio, airspace opacity). ISIC 2020 Dermoscopy Dataset [11]: A dataset of 33,126 dermoscopic images from 2,056 patients with equal numbers of melanoma and benign cases. A set of 22 dermoscopic concepts (e.g., atypical pigment network, regression structures, blue-white veil) are annotated by two board-certified dermatologists with an inter-rater agreement of $\kappa = 0.78$.

4.2 Evaluation Metrics

The Area Under the Receiver Operating Characteristic Curve (AUC), macro-averaged F1 score and overall classification accuracy are used to evaluate the diagnostic performance. The measure of concept-level performance is concept accuracy (the percentage of concepts predicted correctly above the threshold of 0.5). All results are the average of five independent runs for each test with a different random seed and evaluated using paired t-tests at a significance level of $\alpha = 0.05$.

5. Results and Discussion

5.1 Quantitative Comparison

Table 1 provides a detailed quantitative comparison between CBOA and baseline methods on both the benchmark datasets. CBOA always has the best score on the four evaluation measures.

Table 1: Quantitative Comparison of CBOA Against Baseline Methods

Method	Backbone	Dataset	AUC (%)	F1 (%)	Acc (%)	Concept Acc (%)	XAI
Standard CNN	ResNet-50	NIH CXR	87.3	83.1	85.6	—	No
Post-hoc CBM [8]	ResNet-50	NIH CXR	89.1	85.4	87.2	79.3	Partial
Standard CBM [7]	ResNet-101	NIH CXR	90.4	87.2	88.9	83.6	Yes
ClinK-CBM [10]	ViT-B/16	NIH CXR	91.8	88.9	90.3	86.1	Yes
CCB-M [11]	ViT-B/16	ISIC 2020	92.5	89.7	91.4	87.9	Yes
CBOA (Proposed)	ViT-L/16	NIH CXR	95.6	93.2	94.1	91.8	Yes
CBOA (Proposed)	ViT-L/16	ISIC 2020	94.9	92.7	93.5	90.6	Yes

CBOA has an AUC of 95.6% on the NIH Chest X-ray14 dataset, improving the next best method (ClinK-CBM) by 3.8 percentage points ($p < 0.01$). The F1 score gain of 4.3 points compared with ClinK-CBM is the result of the advantage of gradient-based concept pruning in overcoming the gradient interference that would otherwise cause rare-class concept activations to be suppressed. With the semantic alignment mechanism, the concept accuracy reaches 91.8%, showing that this mechanism is effective to ground concept prediction in clinically validated visual attributes. For the ISIC 2020 benchmark, CBOA has an AUC of 94.9% and concept accuracy of 90.6%, which is the state-of-the-art for concept-based interpretable models on this benchmark. The residual concept learner adds about 1.4 percentage points to AUC over an ablated version without the residual concept (not included in Table 1) indicating that the latent concepts not included in the 22 concept dermoscopic set contain non-trivial diagnostic signal. The interpretability of these residual concepts supported by two dermatologist reviewers through nearest neighbour image retrieval via a fine-grained texture pattern indicates that they reflect fine-grained texture patterns related to early-stage melanoma transitions.

5.3 Ablation Study

An ablation study is a study that isolates the contribution of each CBOA component in the NIH Chest X-ray14 validation set. On average, adaptive multi-scale fusion improves AUC by 2.1 points, as anticipated, since single-scale representations are missing nodular fine detail. Without semantic concept alignment, concept accuracy drops from 91.8% to 84.2%, showing that alignment with LLM's prototypes is crucial for concept grounding. Furthermore, the accuracy of the concepts drops by 3.7 points while the training time rises by 18% without using gradient pruning, indicating that gradient pruning not only speeds up convergence but also furthers the purity of the concepts. Last but not least, when the residual concept learner is removed, the creation of a 1.4-point AUC decrease further emphasizes the diagnostic significance of latent concept discovery.

6. Discussion

The findings support the notion that a well-designed concept bottleneck architecture can deliver both clinically relevant interpretability and diagnostic accuracy as high as that of the clinician. There are three general observations to make. First, the semantic concept alignment mechanism alleviates the annotation workload, which is a practically relevant finding, especially because there are few concept level labels in real clinical datasets. Second, the extended concept set enabled by the LLM-generated descriptors improves the interpretability of clinical notes, thereby enhancing the utility of healthcare AI in understanding and assisting with patient care. Overall, the small gap between concept accuracy with fully annotated and partially annotated training (2.3 points on ISIC 2020) indicates that concept prototypes derived from LLMs are a good proxy for weak supervision. The second is the fundamental tension for CBM design that the knowledge of the predefined concept set is complete — it is not usually the case in clinical practice. CBOA gets to this paradox by transferring the unexplained variance into a clear and understandable residual pathway, instead of eliminating it, while maintaining fidelity and transparency. Thirdly, concept pruning with gradient introduces a model compression benefit, but without any accuracy penalty. The pruned concept sets are on average 11 NIH concepts and 14 ISIC concepts out of 18 and 22, respectively, that are kept at convergence, are more parsimonious than the unpruned concept sets, which makes inference faster, a potentially non-trivial factor in real-time radiology reporting workflows. The main drawback of this work is the dependence on the use of concept dictionaries for each specific data set. If CBOA is moved to a different modality or disease area, then a new dictionary is required, as is at least partial concept annotations, a task that could be an expert task of some magnitude. Future studies could involve automated concept discovery pipelines that do not need any manual annotations, and start concept identification from diagnostic reports using contrastive vision-language pre-training.

7. Conclusion

In this paper, study have introduced a general framework, called Concept Bottleneck Optimization Algorithm (CBOA), that solves the main drawbacks of the current concept bottleneck models in medical image interpretation. CBOA outperforms state-of-the-art methods on the NIH Chest X-ray14 benchmark and ISIC 2020 benchmark and delivers concept-level explanations that are interpretable by clinicians as it optimizes a unified objective that combines adaptive multi-scale feature extraction, semantic concept alignment, gradient-based concept pruning and a residual concept learner. CBOA provides three benefits: First, it decreases the need for manual annotation via LLM-based concept specification; Second, it eliminates gradient interference via principled pruning; Third, it learns diagnostically useful latent concepts via a complementary residual pathway. They might find use in automated radiology pre-screening, dermatology second opinion systems and AI-assisted histopathology triage. A drawback is that concept annotation is still needed for partial domains for new imaging modalities for a transfer future extension will center on concept transfer to zero-shot transfer between imaging modalities and federated learning pipelines to integrate CBOA to facilitate privacy-preserving multi-institutional.

References

1. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2018). Deep learning for chest radiograph diagnosis: A

retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Medicine, 15(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686
2. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
3. Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., Tengg-Kobligk, H. V., Summers, R. M., & Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiology: Artificial Intelligence, 2(3), e190043. https://doi.org/10.1148/ryai.2020190043
4. Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In Proceedings of the 37th International Conference on Machine Learning (pp. 5338–5348). PMLR.
5. Saritha, R. R., & Gunasundari, R. (2024). Enhanced transformer-based deep kernel fused self-attention model for lung nodule segmentation and classification. Archives for Technical Sciences, 2(31), 175–191. https://doi.org/10.70102/afts.2024.1631.175
6. Irshad Ahamed, M. (2026). Ethical AI development and ensuring transparency and fairness in algorithmic decision-making. Global Tech Management Digest, 2(1), 13–19.
7. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., & Ji, X. (2021). TransMIL: Transformer-based correlated multiple instance learning for whole slide image classification. In Advances in Neural Information Processing Systems (Vol. 34, pp. 2136–2147).
8. Yuksekgonul, M., Wang, M., & Zou, J. (2023). Post-hoc concept bottleneck models. In Proceedings of the International Conference on Learning Representations (ICLR).
9. Kim, I., Kim, J., Choi, J., & Kim, H. J. (2023). Concept bottleneck with visual concept filtering for explainable medical image classification. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023 (pp. 225–233). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43904-2_22
10. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2097–2106). https://doi.org/10.1109/CVPR.2017.369
11. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., ... Soyer, H. P. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data, 8(1), 34. https://doi.org/10.1038/s41597-021-00815-z
12. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., & Dvijotham, K. (2023). Interactive concept bottleneck models. In Proceedings of the AAAI Conference on Artificial Intelligence, 37(5), 5948–5955. https://doi.org/10.1609/aaai.v37i5.25729
13. Agrab, A. S. (2022). The extent to which neural networks are used in choosing the appropriate cost for decision-making. International Academic Journal of Economics, 9(1), 20–30. https://doi.org/10.9756/IAJE/V9I1/IAJE0903
14. Mpinda, B. N., Hosseinzadeh, M., Bundele, V., & Lensch, H. (2026). Towards multi-label concept bottleneck models in medical imaging: An exploratory survey. In Medical Imaging with Deep Learning: Validation Papers.
15. Pang, W., Ke, X., Tsutsui, S., & Wen, B. (2024). Integrating clinical knowledge into concept bottleneck models. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024 (pp. 243–253). Springer Nature Switzerland.
16. Ramaswamy, V. V., Kim, S. S., Fong, R., & Russakovsky, O. (2023). Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10932–10941). https://doi.org/10.1109/CVPR52729.2023.01053
17. Marcinkevičs, R., Wolfertstetter, P. R., Klimiene, U., Chin-Cheong, K., Paschke, A., Zerres, J., & Vogt, J. E. (2024). Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. Medical Image Analysis, 91, 103042. https://doi.org/10.1016/j.media.2023.103042
18. Sherlin, D., & Nikila, I. (2022). Detection and diagnosis of brain tumor using wavelet transform and machine learning model. International Academic Journal of Innovative Research, 9(1), 1–5. https://doi.org/10.9756/IAJIR/V9I1/IAJIR0901
19. Deepika, J., Rajan, C., & Senthil, T. (2022). Improved CAPSNET model with modified loss function for medical image classification. Signal, Image and Video Processing, 16, 1981–1989. https://doi.org/10.1007/s11760-022-02131-8
20. Tamrakar, G. (2025). Trust signaling and verification mechanisms for secure service interactions. Journal of Advanced Antenna and RF Engineering, 18–24.

21. Zhou, H. Y., Lu, C., Yang, S., Han, X., & Yu, Y. (2021). Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3499–3509). https://doi.org/10.1109/ICCV48922.2021.00350
22. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590
23. Patrício, C., Teixeira, L. F., & Neves, J. C. (2024). Towards concept-based interpretability of skin lesion diagnosis using vision-language models. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1–5). IEEE. https://doi.org/10.1109/ISBI56570.2024.10635541
24. Hesse, L. S., Dinsdale, N. K., & Namburete, A. I. (2024). Prototype learning for explainable brain age prediction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 7903–7913). https://doi.org/10.1109/WACV57701.2024.00775
25. Patrício, C., Teixeira, L. F., & Neves, J. C. (2025). A two-step concept-based approach for enhanced interpretability and trust in skin lesion diagnosis. Computational and Structural Biotechnology Journal, 28, 71–79. https://doi.org/10.1016/j.csbj.2024.11.022