A Controlled Study of Ensemble Classification and Multimodal Explanations for Skin-Lesion Recognition

Žan Kogovšek[1], **Mohd Faizan Ansari**[2], **Žiga Emeršič**[1]

[1]*University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, Ljubljana, Slovenia*
[2]*Department of Applied Informatics, Silesian University of Technology, 44-100 Gliwice, Poland*
zan.kogovsek@fri.uni-lj.si

Abstract

Deep neural networks achieve strong skin-lesion recognition performance, but their use in clinical practice also depends on whether their decisions can be explained reliably. We evaluate 37 convolutional and transformer model variants, analyze the influence of key training choices, and combine the strongest architecturally distinct models into a feature-level ensemble. The ensemble achieves a balanced multiclass accuracy of 84.6% *on the ISIC 2018 test set. We then assess three explanation modalities: class activation visualizations, example-based retrieval, and textual explanations generated by large language models. The best class activation mapping method reaches a Jaccard index of* 0.445 *against lesion segmentation masks, while image retrieval achieves a precision of* 0.702 *among the five nearest examples. Finally, we compare textual explanations generated by four large language models from the dermatoscopic image alone and with an additional class activation visualization. An exploratory dermatologist evaluation indicates that the explanations can contain clinically relevant observations, but may also introduce unsupported, unnecessary, or confusing details. The additional visual input was useful in some cases, but did not provide consistently helpful grounding. These results show that predictive performance and explanation quality must be evaluated separately, since accurate predictions and additional explanatory context do not necessarily produce clinically useful explanations.*

1 Introduction

Early recognition of malignant skin lesions is important for effective clinical management, while dermatoscopic assessment requires specialist expertise and considerable clinical resources [1]. The growing availability of digital dermatoscopic images and advances in deep learning have therefore motivated computer-aided systems intended to support, rather than replace, clinical decision making [2]. Convolutional neural networks and vision transformers have achieved strong performance in medical image analysis, with selected studies reporting results comparable to those of experienced dermatologists [3, 4].

Predictive accuracy alone is insufficient for use in clinical practice. Deep models remain difficult to inspect, and a user may receive a confident prediction without knowing which image evidence supported it or whether the prediction is clinically plausible [5]. Explainable artificial intelligence attempts to address this limitation through visual explanations, retrieval of similar cases, and textual explanations [6]. These approaches provide complementary information: class activation maps indicate *where* a model focused, example-based explanations identify *which previous cases* are represented similarly, and textual explanations describe image characteristics associated with the predicted diagnosis.

This paper studies predictive performance and explanation quality within a single skin-lesion recognition pipeline. The main contributions are:

1. a systematic evaluation of convolutional and transformer architectures and the training choices that influence their performance;
2. a feature-level ensemble combining three strong models from distinct architectural families; and
3. quantitative evaluation of class activation visualizations and example-based retrieval, together with an exploratory dermatologist evaluation of textual explanations with and without additional visual grounding.

The study evaluates a competitively performing ensemble and assesses complementary explanation modalities in terms of spatial alignment, diagnostic relevance, and practical usefulness.

2 Related Work

Skin lesion recognition and ensembles. Pretrained convolutional neural networks have become a common approach to medical image classification, with transfer learning being particularly useful when only limited labeled medical data are available [7]. On ISIC benchmarks, data augmentation and class-imbalance treatment are commonly used because the diagnostic categories are highly skewed [8]. Larger architectures do not necessarily perform better, as their increased capacity may lead to overfitting. Ensembles can improve robustness by combining models with complementary representations and prediction errors [9]. However, reported gains are often difficult to compare because preprocessing, model tuning, data splits, and external training data differ between studies. This motivates a controlled evaluation in which all candidate models share the same training and evaluation protocol.

Explainability in medical imaging. Class activation visualizations are commonly used to examine whether a classifier focuses on the lesion rather than on surrounding tissue, acquisition artifacts, or image background. Example-based explanations provide a complementary form of evidence by retrieving previously observed cases with similar learned representa-

tions [10]. However, a systematic review of explainable skin lesion recognition found that explanations are frequently presented only as qualitative examples and are rarely compared using standardized quantitative evaluation [11]. User studies further suggest that explanations may improve confidence and diagnostic performance, although their usefulness depends on the task, explanation method, and intended user [12]. Textual explanations generated by large language models may present model outputs in a more accessible form, but fluent explanations can include unsupported visual observations or clinically questionable reasoning. Class activation visualizations could provide additional grounding by indicating image regions considered relevant by the diagnostic model. Nevertheless, such visualizations are themselves imperfect explanations. We therefore examine how this additional visual input affects the generated textual explanations rather than assuming that more context is necessarily beneficial.

3 Methods

Dataset and evaluation protocol. Experiments are conducted on the ISIC 2018 dataset, which contains 10,015 training images, 193 validation images, and 1,512 test images from seven diagnostic categories [13, 14]. The official validation split is too small to provide a reliable estimate for all classes, particularly given the strongly imbalanced class distribution. Rather than relying on this split for model selection, we apply five-fold cross-validation to the training set. The folds are stratified by diagnostic class at the image level. This allows every training image to be used for validation once and reduces dependence on a single train–validation split. Balanced multiclass accuracy is used as the recognition metric throughout the study. Images are resized to 224×224 pixels. The models are trained for 70 epochs using Adam with a batch size of 32. All experiments use deterministic operations and a fixed random seed of 2025. Model checkpoints are selected according to validation balanced multiclass accuracy.

Hyperparameter selection. Training choices are evaluated using a greedy sequential search on EfficientNet-B2. At each stage, candidate values are compared while the previously selected settings remain fixed. This procedure is computationally efficient but is not a full ablation. The evaluated choices include the loss function, class weighting and sampling, training and test-time augmentation, learning-rate scheduling, and weight decay. The loss functions include cross-entropy, focal, label-distribution-aware margin, and label-occurrence-weighted loss [15, 16]. Class-imbalance strategies include direct and deferred weighting, where deferred weighting introduces class weights only after an initial unweighted training period. The resulting training configuration is applied uniformly across architecture families to support a controlled comparison. Only the learning rate is selected separately for each architecture. The benchmark therefore prioritizes comparability under a common training protocol rather than architecture-specific hyperparameter optimization.

A learning-rate finder is applied separately to each architecture by exponentially increasing the learning rate during a short training run and selecting a value near the steepest decrease in loss [17]. This accounts for differences in the optimization behaviour of the evaluated architectures. Because the selected learning rate is incorporated into the architecture benchmark, its individual contribution is not treated as a separate ablation.

Training augmentation compares modified RandAugment with custom strategies that combine colour and shape changing transformations, collectively referred to as the Dual-Category augmentation strategy. At inference, the Multiple Crops strategy divides each image into 16 overlapping regions, obtains an independent prediction for each crop, and averages the resulting outputs. The procedure is illustrated in Figure 1.

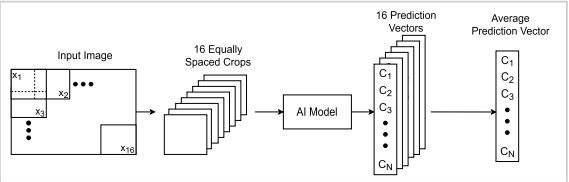


Figure 1: Multiple-crop test-time augmentation. Sixteen overlapping crops are classified independently and their predictions are averaged.

Architecture benchmark and ensemble construction. All models are initialized with ImageNet-pretrained weights and fine-tuned for skin-lesion recognition. We evaluate 37 variants from nine convolutional and transformer families: EfficientNet, EfficientNet V2, ResNet, ResNeXt, DenseNet, ConvNeXt, Vision Transformer, Swin Transformer, and Swin Transformer V2. Individual candidate models are compared using cross-validation, while the test set is reserved for evaluation of the final ensemble.

The ensemble combines EfficientNet-V2-L, ConvNeXt-Large, and Swin-Base, which provide strong validation performance and belong to distinct architectural families. Their parameters are frozen, their feature representations are concatenated, and trainable linear layers map the combined representation to the seven diagnostic classes.

The same five-fold partitions are used throughout the ensemble evaluation. Within each fold, the constituent-model checkpoints are selected according to validation balanced multiclass accuracy. The fusion head is trained using the same training partition as the corresponding constituent models, while the held-out validation partition is used to select the fusion-head checkpoint. This produces five fold-specific ensembles. The final test performance is reported as the mean and standard deviation across these five ensembles. For comparison with leaderboard entries, which report one score per submission, we additionally report the test performance of the ensemble fold selected by its validation performance.

Class activation visualizations. Class activation maps are generated from the final non-pooling layer of each ensemble member. For each input image, the predicted class is used as the explanation target, and the maps produced by the three models are averaged into a single ensemble visualization. We compare GradCAM, GradCAM++, EigenCAM, XGradCAM, and LayerCAM [18, 19, 20, 21, 22].

Each class activation map is normalized to the interval $[0, 1]$ and converted into a binary mask by marking pixels above a selected threshold as activated. Thresholds from 0.1 to 0.9 are evaluated using GradCAM on the training-set images. To keep the binarization procedure fixed across methods, the selected threshold is then applied to all five class activation mapping methods. This provides a controlled comparison without independently tuning each method, although differences in their activation-score distributions mean that the shared threshold may not be optimal for every method. The resulting masks

are compared with the lesion segmentation masks using the Jaccard index:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{1}$$

where A denotes the activated image region and B the ground-truth lesion region. This evaluation measures whether the regions emphasized by the visualization overlap with the lesion, but does not determine whether these regions alone are responsible for the prediction.

Example-based explanations. Example-based explanations are generated through content-based image retrieval. A query representation $\mathbf{f}_q$ is extracted from the ensemble's concatenated feature layer, and training images are ranked according to their Euclidean distance from the query:

$$d(\mathbf{f}_q, \mathbf{f}_i) = \sqrt{\sum_{j=1}^{n} (f_{qj} - f_{ij})^2}. \tag{2}$$

A retrieved image is considered relevant when its ground-truth diagnostic class matches that of the query. Retrieval performance is evaluated using precision, recall, and F1 score for $K \in \{5, 10, 20\}$. This evaluates the diagnostic consistency of the learned representation, although images from the same class are not necessarily morphologically identical.

Exploratory evaluation of textual explanations. We conduct an exploratory evaluation of textual explanations generated by GPT-5.5, Claude 4.6 Sonnet, Grok 4.1, and Gemini 3.1 Pro. Each model receives the same zero-shot prompt under two conditions. In the baseline condition, the input consists of the RGB image and the predicted diagnostic class. In the visually grounded condition, a GradCAM visualization is additionally provided as supportive attention evidence. This experiment was conducted separately from the quantitative comparison of class activation mapping methods, in which GradCAM++ achieved the highest lesion-mask overlap.

Three correctly classified ISIC 2018 cases are evaluated: melanoma, nevus, and benign keratosis. The generated explanations are blinded, randomized, and assigned neutral identifiers. One specialized dermatologist ranks the four explanations from 1, representing the most clinically accurate explanation, to 4, representing the least accurate, within each case and input condition. Written feedback addresses correctly identified properties, unsupported claims, use of the class activation visualization, trust, patient communication, and possible integration into clinical practice. Because the evaluation contains three cases and one dermatologist, the rankings are interpreted descriptively rather than as evidence of model superiority or improvement between input conditions.

4 Results

Hyperparameter selection. Figure 2 summarizes the incremental hyperparameter experiments. Starting from a validation balanced multiclass accuracy of 79.837%, the selected configuration increased the score to 89.228%. The final configuration used label-distribution-aware margin loss, balanced class weights with deferred weighting, Dual-Category training augmentation, Multiple Crops at inference, a `ReduceLROnPlateau` scheduler, and no weight decay. The largest improvement was obtained when Dual-Category augmentation was combined with Multiple Crops, increasing the score from 82.485% to 88.871%.

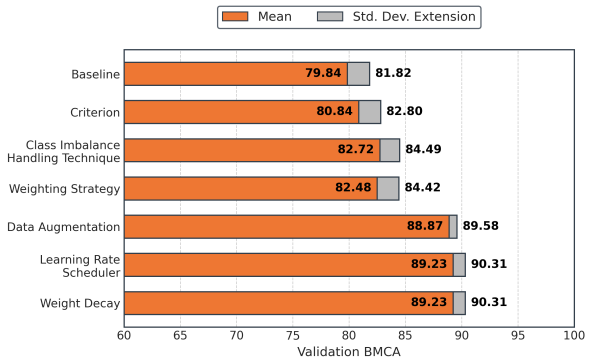


Figure 2: Incremental validation performance as the selected training choices are introduced and fixed.

Architecture and ensemble performance. Table 1 presents the strongest variant from each evaluated model family. ConvNeXt-Large achieved the highest individual validation score, followed by EfficientNet-V2-L and Swin-Base. These three models were selected for the ensemble because they combine strong individual performance with architectural diversity.

Table 1: Five-fold mean validation balanced multiclass accuracy ($\pm$ standard deviation) of the strongest model from each evaluated family.

Model	Validation BMCA (%)
EfficientNet-B2	89.902 ± 0.710
EfficientNet-V2-L	90.172 ± 1.056
ResNet-152	86.018 ± 1.432
ResNeXt-101-64x4d	89.040 ± 1.210
DenseNet-161	88.165 ± 1.191
ConvNeXt-Large	90.889 ± 1.346
ViT-L-16	88.914 ± 1.574
Swin-Base	90.128 ± 0.910
Swin-V2-Small	89.751 ± 1.675

The test set was not used to report or select individual constituent models and was used only for final evaluation. The five fold-specific ensembles achieved a mean test balanced multiclass accuracy of $84.624 \pm 0.951\%$, where the standard deviation is calculated across their five test scores and reflects variation due to the different training folds. The ensemble fold selected solely according to validation performance achieved a test score of 85.1%; this single score is used only for comparison with the single-score leaderboard submissions in Table 2.

For comparison, Table 2 distinguishes between results submitted during the original ISIC 2018 challenge and the best result subsequently reported on the live leaderboard. Because leaderboard submissions report a single test score rather than a cross-validation mean, the proposed row uses the test performance of the ensemble fold selected according to validation performance. Its balanced multiclass accuracy of 85.1% exceeds the selected original challenge submissions, including the 84.5% result of Sun Yat-sen University, but remains below the later live-leaderboard result of 88.6% [13].

Class activation visualizations. Thresholds from 0.1 to 0.9 were evaluated using GradCAM on the training-set images, with 0.3 producing the highest Jaccard index. This shared threshold was then used to compare the five class activation

Table 2: Single-score comparison with selected original ISIC 2018 challenge submissions and the later live-leaderboard result.

Source	Team or approach	BMCA (%)
Original challenge	Ask Sina	81.2
Original challenge	Li	81.5
Original challenge	Sun Yat-sen University	84.5
Proposed method	*Proposed ensemble*	*85.1*
Live leaderboard	**Best reported result**	**88.6**

mapping methods reported in Table 3. GradCAM++ achieved the highest Jaccard index and was therefore selected for evaluation on the independent test set, where it obtained a score of 0.445.

Table 3: Jaccard index of the evaluated class activation mapping methods using the shared threshold of 0.3 selected with GradCAM. Only GradCAM++ was subsequently evaluated on the test set.

Method	Training	Test
GradCAM	0.380	–
GradCAM++	**0.411**	**0.445**
EigenCAM	0.374	–
XGradCAM	0.379	–
LayerCAM	0.389	–

Example-based explanations. The retrieval results are reported in Table 4. Precision was highest when five examples were retrieved and decreased as the number of retrieved examples increased. The low recall values reflect the evaluation definition: each query may have many relevant same-class images in the training set, while only a small number of examples are returned. For an explanation interface intended to display a concise set of similar cases, precision at small values of K is therefore the more informative measure. Figure 3 illustrates an example explanation consisting of a query lesion and its five nearest retrieved cases.

Table 4: Example-based explanation results on ISIC 2018.

K	Precision	Recall	F1
5	0.702	0.009	0.017
10	0.673	0.013	0.025
20	0.535	0.019	0.035

Exploratory evaluation of textual explanations. Table 5 summarizes the mean ranks assigned within each input condition. Because the study uses forced rankings across only three cases, the values are treated descriptively and are not used to infer that one language model is generally superior or that visual grounding improves an individual model's explanations. Gemini received the best mean rank in the visually grounded condition, while Gemini and Claude shared the best mean rank in the RGB-only condition.

The written feedback showed that the stronger explanations identified relevant properties such as border and network regularity and proposed appropriate differential diagnoses. Nevertheless, fluent explanations could include observations unsupported by the available input, suggesting possible hallucinations. In one case, an explanation referred to the "ugly duckling" sign, even though this criterion requires comparison with

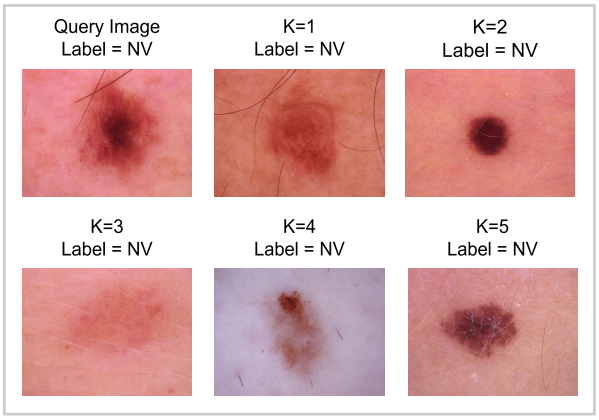


Figure 3: Example-based explanation for a nevus query. The query lesion is shown together with the five nearest training images in the ensemble feature space.

Table 5: Descriptive mean dermatologist ranks across three cases within each input condition (1 = best, 4 = worst).

Model	RGB	RGB + GradCAM
ChatGPT	3.33	3.67
Claude	1.67	2.00
Grok	3.00	3.33
Gemini	1.67	1.00

the patient's other lesions and only one cropped image was provided.

The class activation visualization was also not consistently useful. It helped identify suspicious regions in the melanoma case, but appeared less coherent for benign lesions and sometimes increased verbosity without clarifying the central explanation. The dermatologist therefore viewed the system primarily as a supplementary second opinion rather than an alternative to established diagnostic procedures. Textual explanations were considered particularly useful for patient communication when a clinically alarming-looking lesion, such as a highly pigmented benign keratosis, is assessed as benign but causes patient concern. In such cases, a concise explanation of the visible features supporting the benign assessment could help the dermatologist justify the decision and reduce patient anxiety. However, the explanation should be verified by the clinician and adapted to the patient by removing unnecessary technical terminology and unsupported claims.

5 Discussion

Predictive performance. The incremental experiments show that the final performance results from several interacting training choices rather than from model architecture alone. Loss selection and class weighting provide moderate improvements, while the largest gain is obtained by combining Dual-Category training augmentation with Multiple Crops at inference. Although constant weighting achieved a marginally higher intermediate validation score, deferred weighting was retained to preserve the training formulation proposed together with label-distribution-aware margin loss. The final configuration therefore reflects both empirical performance and methodological consistency rather than selecting every component solely by its highest observed score. Nonzero weight decay does not improve performance.

The architecture comparison shows similar performance

among the strongest convolutional and transformer models. ConvNeXt-Large achieves the highest individual validation score, while EfficientNet-V2-L and Swin-Base provide comparably strong but architecturally distinct representations. Their combination therefore offers a suitable basis for an architecturally diverse ensemble. The resulting ensemble achieves competitive performance on the ISIC 2018 test set.

Explanation quality. The quantitative results suggest that the ensemble provides potentially informative visual and example-based explanations. GradCAM++ achieves the strongest spatial agreement with the lesion segmentation masks, indicating that the highlighted regions are generally concentrated on the lesion. However, the moderate Jaccard score also shows that the visualization does not precisely follow the complete lesion area. Because the comparison uses a single GradCAM-selected threshold for all methods, part of the observed ranking may reflect differences in activation-score distributions rather than localization quality alone. For example-based explanations, the precision of 0.702 at $K = 5$ indicates that most of the nearest retrieved cases share the query diagnosis. The decrease in precision at larger retrieval depths suggests that presenting a small number of examples is more appropriate for a practical explanation interface.

The dermatologist feedback shows that textual explanations can identify relevant dermatoscopic properties and provide appropriate differential diagnoses, but can also contain unsupported observations, suggesting possible hallucinations. Adding a class activation visualization did not have a consistently positive effect: it provided useful attention cues in some cases, particularly for melanoma, but was less informative for benign lesions and sometimes made the explanation more verbose or less focused. Visual, example-based, and textual explanations should therefore be presented as complementary sources of information whose outputs may agree or reveal different aspects of the prediction, rather than being combined into a single authoritative explanation.

6 Conclusion

This study jointly evaluates predictive performance and explanation quality in a skin-lesion recognition pipeline. Across 37 convolutional and transformer variants, both model architecture and training choices substantially influence recognition performance. An ensemble combining EfficientNet-V2-L, ConvNeXt-Large, and Swin-Base achieves a test balanced multiclass accuracy of $84.624 \pm 0.951\%$ on ISIC 2018.

GradCAM++ provides the strongest spatial agreement with lesion segmentation masks, while content-based image retrieval returns diagnostically consistent examples at small retrieval depths. The exploratory dermatologist evaluation shows that textual explanations can identify clinically relevant properties, but may also contain unsupported claims. Providing a GradCAM visualization as additional input was useful in some cases, but did not consistently produce clearer or more clinically appropriate explanations.

Overall, predictive performance, spatial alignment, retrieval consistency, and textual plausibility should be evaluated separately. The study is limited by its use of a single dataset, greedy hyperparameter selection, the label-based definition of retrieval relevance, the use of one shared threshold for class activation mapping methods, and a textual-explanation evaluation involving three cases and one dermatologist. Future work should evaluate retrieved examples according to morphological similarity and validate textual explanations using more cases, multiple clinical experts, and both correct and incorrect ensemble predictions.

References

[1] Anthony F Jerant, Jennifer T Johnson, Catherine Demastes Sheridan, and Timothy J Caffrey. Early detection and treatment of skin cancer. *American family physician*, 62(2):357–368, 2000.

[2] Dhivya Sudeep. Augmented ai in health diagnostics: Enhancing medical decision making through artificial intelligence.

[3] Dmitriy Gavrilov, Alexander Melerzanov, Nikolay Schelkunov, and Artem Gorodilov. Artificial intelligence image recognition inhealthcare. In *2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI)*, pages 24–26. IEEE, 2018.

[4] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, 2019.

[5] Fei Wang, Rainu Kaushal, and Dhruv Khullar. Should health care demand interpretable artificial intelligence or accept "black box" medicine?, 2020.

[6] Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

[7] Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022.

[8] Peng Yao, Shuwei Shen, Mengjuan Xu, Peng Liu, Fan Zhang, Jinyu Xing, Pengfei Shao, Benjamin Kaffenberger, and Ronald X Xu. Single model deep learning on imbalanced small datasets for skin lesion classification. *IEEE transactions on medical imaging*, 41(5):1242–1254, 2021.

[9] Su Myat Thwin and Hyun-Seok Park. Skin lesion classification using a deep ensemble model. *Applied Sciences*, 14(13):5599, 2024.

[10] Stefano Allegretti, Federico Bolelli, Federico Pollastri, Sabrina Longhitano, Giovanni Pellacani, and Costantino Grana. Supporting skin lesion diagnosis with content-based image retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8053–8060. IEEE, 2021.

[11] Katja Hauser, Alexander Kurz, Sarah Haggenmueller, Roman C Maron, Christof von Kalle, Jochen S Utikal, Friedegund Meier, Sarah Hobelsberger, Frank F Gellrich, Mildred Sergon, et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer*, 167:54–69, 2022.

[12] Mahya Sadeghi, Parmit Chilana, Jordan Yap, Philipp Tschandl, and M Stella Atkins. Using content-based image retrieval of dermoscopic images for interpretation and education: A pilot study. *Skin Research and Technology*, 26(4):503–512, 2020.

[13] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). February 2019.

[14] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5(1):180161, August 2018.

[15] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[16] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[17] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[18] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[19] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRR*, abs/1710.11063, 2017.

[20] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *CoRR*, abs/2008.00299, 2020.

[21] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *CoRR*, abs/2008.02312, 2020.

[22] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE transactions on image processing*, 30:5875–5888, 2021.