

A Graph Path Interpretability Metric for Explainable Classification of Dermoscopic Skin Lesions

Aleksandra N. Zhdanova

Samara National Research University, 34 Moskovskoe Shosse, Samara, 443086, Russian Federation

e-mail: aleksandra.zhd@yandex.ru

Abstract. Deep neural networks can classify dermoscopic images of skin lesions at a level comparable to that of medical experts. However, their lack of transparency limits the interpretability of model decisions, while existing explainable artificial intelligence techniques often produce pixel-level attributions that are difficult to relate to clinical diagnostic criteria. This study proposes the Graph Path Interpretability (GPI) metric, a quantitative model-independent framework for assessing whether a model decision can be explained through a clinically meaningful path of the form “image – dermoscopic feature – diagnosis”. The study was conducted using the HAM10000 dataset containing 10015 images from seven diagnostic classes. An EfficientNet-B0 classifier achieved an accuracy of 0.81 and an AUC of 0.96 and was combined with five ABCD-rule features and a knowledge graph linking dermoscopic features and diagnoses through statistically significant associations identified using the Benjamini–Hochberg correction. Interpretability was evaluated using the path coverage metric GPI_{cov} , the weighted path metric GPI_w , and a faithfulness test based on sequential concept removal. Three reasoning modules – graph attention, graph convolution, and multilayer perceptron (MLP), were compared under similar classification performance. The value of GPI_{cov} was close to 1.0, indicating that nearly all model decisions could be associated with conceptual paths in the knowledge graph. Under comparable predictive performance, graph-based models generated decisions supported by clinically meaningful paths, whereas the MLP model did not exploit conceptual relationships. The concept-removal test was consistent with the generated explanations. Evaluation on the independent ISIC 2018 dataset yielded an accuracy of 0.78 and an AUC of 0.94, while GPI_{cov} remained close to 1.0. The results indicate that GPI can serve as a complementary quantitative measure of interpretability for dermoscopic image classification models.

Keywords: dermoscopy; skin cancer; explainable artificial intelligence; interpretability; knowledge graph; deep learning.

Paper #9639 received 20 Jun 2026; revised manuscript received 20 Jul 2026; accepted for publication 21 Aug 2026; published online 29 Sep 2026. [doi: 10.18287/JBPE26.12.030304](https://doi.org/10.18287/JBPE26.12.030304).

1 Introduction

Dermoscopy is an optical imaging technique that enhances the visualization of subsurface skin structures and substantially improves the early detection of melanoma and other malignant skin lesions compared with unaided visual examination [1]. Over the past

decade, deep convolutional neural networks (CNNs) trained on large dermoscopic image repositories have achieved performance comparable to, and in some cases exceeding, that of board-certified dermatologists in lesion classification tasks [2, 3]. The release of curated datasets such as HAM10000, which contains 10015 dermoscopic images from seven diagnostic

categories [4], has accelerated progress in this field and enabled reproducible evaluation of machine-learning methods.

Machine learning has also become a central component of optical skin cancer diagnostics in biomedical photonics. Convolutional and transformer-based architectures have been applied to the classification of hyperspectral images of skin lesions [5, 6]; dermoscopic images have been combined with spectral information for lesion identification [7]; and Raman and autofluorescence spectroscopy have been used for optical biopsy of melanoma [8].

Despite the high classification performance of modern deep-learning models, interpreting their decisions remains a challenge for clinical deployment. A model that outputs a prediction such as “melanoma, $p = 0.93$ ” without explaining its rationale limits error analysis, clinical decision-making, and practical adoption [9]. To improve transparency, various explainable artificial intelligence (XAI) techniques have been proposed, including saliency maps, Grad-CAM, and SHAP, which identify influential image regions or estimate feature contributions to a prediction [10, 11]. However, these explanations are often difficult to interpret in clinical terms. Saliency and attribution methods highlight important image regions or features but do not directly relate them to diagnostic criteria routinely used in dermoscopy, such as the ABCD rule and dermoscopic structures [12]. In addition, interpretability is commonly assessed qualitatively, making quantitative comparison between models difficult.

This study introduces the Graph Path Interpretability (GPI) metric, which represents an explanation as a path of the form “image – dermoscopic feature – diagnosis”. Dermoscopic features derived from the ABCD rule are used as the conceptual layer, while relationships between features and diagnostic categories are represented by a knowledge graph [13].

The main contributions of the study are as follows:

- A GPI metric for the quantitative assessment of interpretability in dermoscopic image classification models is proposed.
- A knowledge graph linking ABCD-rule features to the diagnostic categories of the HAM10000 dataset is developed.
- Graph-based and non-graph-based models are compared under similar classification performance, and their interpretability is evaluated quantitatively.

2 Materials and Methods

2.1 Dataset and Data Splitting

The HAM10000 dataset [4] was used in this study. It contains 10015 dermoscopic images from seven diagnostic categories: melanocytic nevi (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses and intraepithelial carcinoma (akiec), vascular lesions (vasc), and

dermatofibroma (df). The class distribution is highly imbalanced, with melanocytic nevi representing the largest category. Because images may correspond to the same lesion, the dataset was split into training, validation, and test subsets using the lesion identifier (lesion_id). This prevented images of the same lesion from appearing in different subsets, avoiding information leakage between training and evaluation data.

The dataset was divided into training, validation, and test sets in an approximate ratio of 70:15:15, corresponding to 7008, 1515, and 1492 images, respectively. To ensure reproducibility of the experiments, a fixed random seed (seed = 42) was used throughout all data-splitting procedures.

2.2 Baseline Classification Model

EfficientNet-B0 [14], pretrained on the ImageNet dataset, was used as the baseline classification model and subsequently fine-tuned to classify dermoscopic images into seven diagnostic categories. All images were resized to 224×224 pixels.

To address class imbalance, a class-weighted loss was employed, with weights inversely proportional to class frequencies. Model training used the AdamW optimizer (learning rate = 1×10^{-4} , weight decay = 1×10^{-4}) for 30 epochs. The final model was selected according to the highest validation accuracy. Classification performance was evaluated using Accuracy, Balanced Accuracy, macro-F1, and macro-AUC. Ninety-five-percent confidence intervals for all test-set metrics were obtained by non-parametric bootstrap resampling of the test set (2000 resamples).

2.3 Dermoscopic Features

Five interpretable features corresponding to the ABCD rule [12] were extracted for each image. All features were normalized to the range [0, 1]:

- Asymmetry (A) – the degree of mismatch between the lesion mask and its reflections across the principal axes passing through the centroid.
- Border irregularity (B) was defined as $1 - \text{circularity}$, where circularity is calculated as $4\pi \times \text{area} / \text{perimeter}^2$.
- Color variation (C) was measured as the number of color clusters obtained using k-means clustering in the CIE-Lab color space and occupying at least 5% of the lesion area.
- Diameter (D) – the equivalent lesion diameter normalized by image size.
- Structure (S) – structural heterogeneity measured as the normalized Shannon entropy of the lesion brightness histogram.

Lesion masks from the HAM10000 segmentation dataset [4] were used whenever available. For images without provided segmentations, masks were generated using Otsu thresholding applied to the inverted grayscale image. Since all features are computed directly from the image and lesion mask, they are independent of the

classification architecture, ensuring the model-independent nature of the proposed GPI metric.

2.4 Knowledge Graph Linking Features and Diagnoses

A knowledge graph linking dermoscopic features and diagnostic categories was constructed. Graph edges represent the strength and statistical significance of associations between feature c and diagnosis k , estimated exclusively on the training set.

Edge weights were computed using the point-biserial correlation coefficient r_{ck} between the feature value and a one-vs-rest binary class indicator. Multiple hypothesis testing was controlled using the Benjamini-Hochberg procedure. Associations were considered statistically significant when $q < 0.05$ and $r > 0$. Only statistically significant feature–diagnosis associations were included in the graph. Out of 35 possible feature–diagnosis pairs, 11 positive associations satisfied the significance criteria and were retained as graph edges.

2.5 Reasoning Modules

To compare interpretability, three reasoning modules were trained using the same set of five dermoscopic features. Classification was performed on the basis of a shared conceptual feature representation [15].

GAT – feature nodes are aggregated using a graph attention mechanism [16]. The embedding produced by the baseline classifier serves as the attention query. The resulting feature weights w_i , normalized with a softmax function, are interpreted as path weights in the computation of GPI.

GCN – feature nodes are aggregated through graph convolution without an attention mechanism. In the GPI calculation, path weights are determined solely by the presence of the corresponding features.

MLP – the five-dimensional feature vector is provided directly to a multilayer perceptron without using graph structure.

All models were trained under identical optimization settings (AdamW, 50 epochs) and selected according to validation-set performance. The use of a common conceptual representation minimizes the influence of classification accuracy differences when comparing model interpretability.

2.6 GPI Metric

For each image, explanation paths of the form “image – feature – diagnosis” are considered. A path is regarded as valid if feature i is present in the image $c_i > \tau$, $\tau = 0.2$) and the association between this feature and the predicted class $\hat{y}_n$ is statistically significant $q < 0.05$.

The path coverage metric quantifies the proportion of predictions for which at least one valid path exists between the image and the predicted class:

$$GPI_{cov} = \frac{1}{N} \sum_{n=1}^N I_n,$$

where $I_n = 1$ if at least one feature i satisfies $c_i^{(n)} > \tau$ and has a significant edge to the predicted class $\hat{y}_n$, and $I_n = 0$ otherwise.

The weighted interpretability metric measures the proportion of the model contribution associated with valid explanation paths:

$$GPI_w = \frac{1}{N} \sum_1^N \frac{\sum_i w_i^{(n)} \cdot s_i^{(n)}}{\sum_i w_i^{(n)}},$$

where, w_i denotes the importance of feature i for the reasoning module, whereas s_i indicates the clinical validity of the corresponding path for the predicted class. Consequently, GPI_w represents the fraction of the model decision weight attributable to clinically valid explanation paths.

For the GAT model, the values w_i correspond to attention coefficients. For the GCN model, w_i are derived from normalized feature values. The metric GPI_w is not computed for the MLP model because it does not produce explicit feature importance weights.

Both metrics are bounded within the interval [0,1]. High values of GPI_{cov} indicate that explanations can be constructed for most model decisions using paths in the knowledge graph. High values of GPI_w indicate that a substantial proportion of the model contribution is associated with valid feature–diagnosis paths.

2.7 Faithfulness Evaluation

To assess the consistency between generated explanations and model behaviour, a sequential concept-removal test was employed [17]. Features were iteratively removed in descending order of their importance w_i , and the resulting change in the confidence of the predicted class was measured.

If the removal of the most important features leads to a substantial decrease in model confidence, the explanation is considered faithful to the underlying decision-making process. Quantitative evaluation was performed using the area under the deletion curve (Deletion AUC), where lower values indicate a stronger influence of the removed features on the classification outcome.

For the MLP model, which does not generate explicit feature importance weights, the feature-removal order was selected randomly and used as a baseline control condition.

3 Results

3.1 Classification Performance

The EfficientNet-B0 model achieved an Accuracy of 0.807, a Balanced Accuracy of 0.677, a macro-F1 score of 0.668, a macro-AUC of 0.955 on the test set (Fig. 1).

The corresponding 95% bootstrap confidence intervals were [0.787, 0.826] for accuracy, [0.622, 0.728] for balanced accuracy, [0.617, 0.708] for macro-F1 and [0.947, 0.962] for macro-AUC; a per-split comparison is given in Table 1.

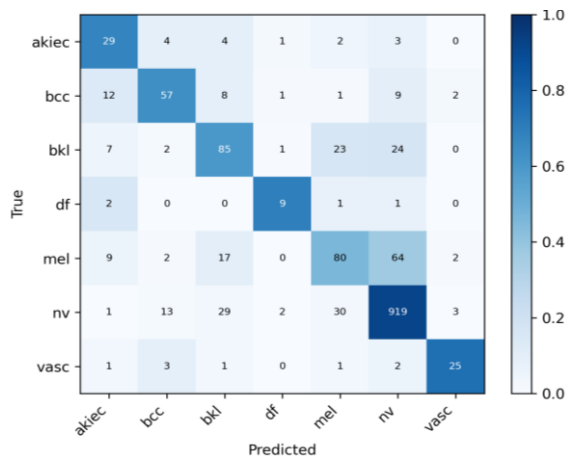


Fig. 1 Normalized confusion matrix for the EfficientNet-B0 classifier on the HAM10000 test set ($n = 1492$). Rows are true diagnoses and columns predicted diagnoses; each row is normalized to sum to one, so the diagonal gives the per-class recall. Class abbreviations: akiec – actinic keratosis, bcc – basal cell carcinoma, bkl – benign keratosis-like lesion, df – dermatofibroma, mel – melanoma, nv – melanocytic nevus, vasc – vascular lesion.

The highest classification performance was obtained for the melanocytic nevus class nv, which represents the largest category in the dataset ($n = 997$ in the test set). The lowest F1 scores were observed for rare classes, particularly dermatofibroma (df, $n=13$), reflecting the pronounced class imbalance of the HAM10000 dataset.

Table 1 Classification performance of the EfficientNet-B0 model on the training, validation and test splits (7008/1515/1492 images). The close agreement between the validation and test metrics, together with the narrow bootstrap confidence intervals, indicates that the reported test performance is a stable estimate; the high training accuracy reflects the expected fit to the training split under early stopping on the validation set.

Split	Accuracy	Balanced Accuracy	macro-F1	macro-AUC
Training	0.996	0.999	0.993	1.000
Validation	0.831	0.692	0.697	0.957
Test	0.807	0.677	0.668	0.955
Test, 95% CI	0.787–0.826	0.622–0.728	0.617–0.708	0.947–0.962

Table 2 Point-biserial correlation coefficients (r) between dermoscopic features (rows) and diagnostic categories (columns), computed on the HAM10000 training set ($n = 7008$). Positive coefficients indicate association with the corresponding diagnosis. Decimal separators are dots.

Feature	akiec	bcc	bkl	df	mel	nv	vasc
Asymmetry	0.02	−0.01	0.00	−0.06	0.03	−0.01	−0.04
Border Irregularity	0.12	−0.01	0.15	−0.06	0.13	−0.19	−0.10
Color Variation	−0.05	−0.19	−0.06	−0.01	0.02	0.14	−0.01
Diameter	0.13	−0.01	0.10	0.03	0.10	−0.17	−0.09
Structural Heterogeneity	−0.09	−0.13	−0.14	−0.10	0.22	0.07	−0.01

Abbreviations: akiec – actinic keratosis, bcc – basal cell carcinoma, bkl – benign keratosis-like lesion, df – dermatofibroma, mel – melanoma, nv – melanocytic nevus, vasc – vascular lesion.

The obtained results are consistent with previously reported performance levels for models of similar complexity on the HAM10000 dataset [4] and were subsequently used as the baseline for interpretability evaluation.

3.2 Knowledge Graph of Feature-Diagnosis Associations

The results of the correlation analysis are summarized in Table 2, and the corresponding knowledge graph is shown in Fig. 2. Of the 35 examined feature-diagnosis pairs, 25 were found to be statistically significant $q < 0.05$, including eleven positive and fourteen negative associations.

Only positive associations $r > 0$, were included in the knowledge graph, resulting in a total of eleven edges.

The largest number of associations were identified for melanoma. This class showed statistically significant positive correlations with structural heterogeneity ($r = 0.22$), border irregularity ($r = 0.13$), lesion diameter ($r = 0.10$), and asymmetry ($r = 0.03$). Positive associations with border irregularity and diameter were also observed for actinic keratosis and benign keratosis-like lesions. No positive associations were identified for basal cell carcinoma or vascular lesions.

The identified relationships are consistent with established clinical knowledge regarding dermoscopic characteristics of skin lesions. In particular, the strongest associations observed for melanoma involved structural heterogeneity and border irregularity.

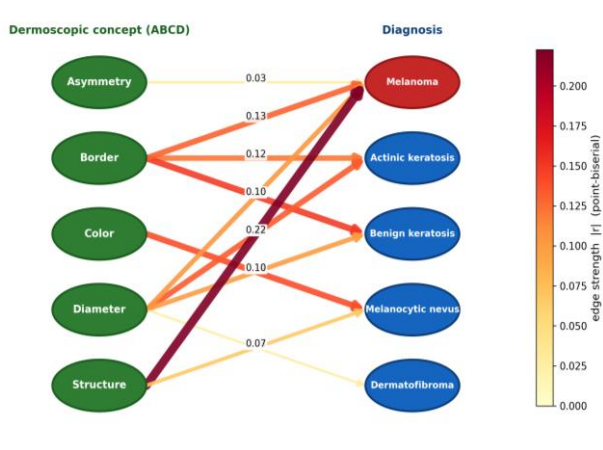


Fig. 2 Knowledge graph of the eleven statistically significant positive feature–diagnosis associations (Benjamini-Hochberg $q < 0.05$, $r > 0$) obtained on the HAM10000 training set. Green nodes are the five ABCD-derived dermoscopic concepts (left) and coloured nodes are diagnoses (right); a directed edge concept $\rightarrow$ diagnosis denotes a significant positive point-biserial correlation, with edge thickness and colour proportional to $|r|$ (colour bar). Basal cell carcinoma (bcc) and vascular lesions (vasc) have no positive edge and are omitted, as their only significant associations are negative (Table 2).

The complete results of the correlation analysis are presented in Table 2. Two observations are noteworthy. First, the absence of positive edges for basal cell carcinoma and vascular lesions does not imply the absence of feature–diagnosis relationships. For these classes, statistically significant negative associations were identified but were excluded from the knowledge graph according to the adopted edge-construction rule. Second, the only positive association involving the color feature was observed for melanocytic nevi $r = 0.14$, whereas no statistically significant association was found

for melanoma. This result may be attributable to the use of the number of color clusters as a measure of color variation and suggests that more detailed color descriptors could improve the representation of chromatic information.

3.3 Comparison of Reasoning Architectures Using GPI

The comparison results for the reasoning modules are presented in Table 3 and Fig. 3. All models used the same set of five conceptual features and differed only in the way these features were aggregated. The Accuracy values for GAT, GCN, and MLP were 0.670, 0.668, and 0.676, respectively, allowing interpretability to be compared under similar classification performance. For all models, the value of GPI_{cov} was close to unity (1.000 for GAT and GCN, and 0.999 for MLP), indicating that at least one knowledge-graph path linking the observed features to the predicted diagnostic class existed for almost all predictions.

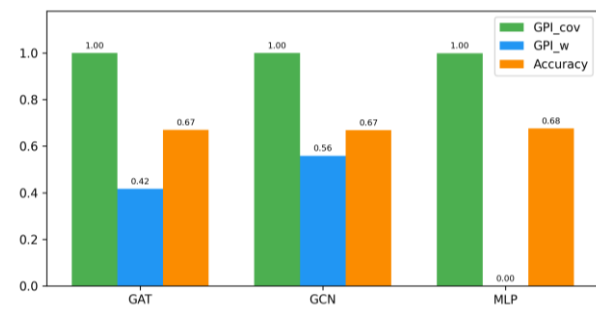


Fig. 3 Interpretability metrics (GPI_{cov} and GPI_w) and classification accuracy of the three reasoning modules (GAT, GCN, MLP). Despite similar accuracy (0.67–0.68), only the graph-based modules achieve non-zero GPI_w , indicating reliance on clinically valid feature–diagnosis paths.

Table 3 Test-set classification accuracy, interpretability metrics (path coverage GPI_{cov} and weighted interpretability GPI_w) and faithfulness (deletion AUC) for the three reasoning modules trained on the same five dermoscopic features of the full HAM10000 dataset (10015 images; test $n = 1492$). A lower deletion AUC denotes a more faithful explanation; for the MLP the value is a random-order control (Fig. 4). The 95% bootstrap confidence intervals, calculated from 2000 resamples of the test set, overlap for all three modules (GAT [0.64, 0.69], GCN [0.65, 0.69], and MLP [0.65, 0.70]), indicating that their accuracies are statistically indistinguishable; for the graph-based modules, $GPI_{cov} = [1.00, 1.00]$. The within-test-set sampling variability of GPI_w is negligible; its dominant source of variation is the training random seed, which is why the GAT-GCN ordering of GPI_w is treated as indicative rather than definitive.

Model	Test Accuracy	GPI_{cov}	GPI_w	Deletion AUC
GAT	0.670	1.000	0.416	0.885
GCN	0.668	1.000	0.559	0.953
MLP	0.676	0.999	0.000	0.691 (Random Baseline)

Abbreviations: GAT – graph attention, GCN – graph convolution, MLP – multilayer-perceptron baseline without graph structure.

The largest differences were observed for GPI_w . Its values were 0.416 for GAT and 0.559 for GCN, whereas GPI_w was equal to 0 for MLP. This result indicates that a substantial fraction of the decision process in graph-based models is associated with clinically meaningful paths represented in the knowledge graph. In contrast, the MLP model does not utilize conceptual relationships and therefore cannot assign decision weight to knowledge-graph paths.

3.4 Faithfulness Evaluation

The results of the sequential feature-removal test are presented in Fig. 4, which shows, for each module, how the predicted-class probability decays as the most important dermoscopic features are progressively removed.

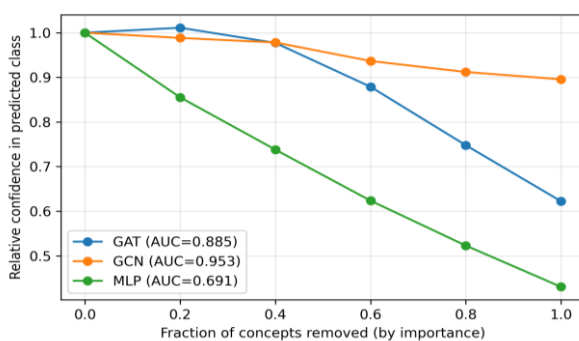


Fig. 4 Faithfulness assessment by sequential concept removal. The five dermoscopic features are deleted one at a time in decreasing order of their model-assigned importance w_i ; the curve shows the retained probability of the predicted class (normalized to its undeleted value, y-axis) versus the fraction of features removed (x-axis). Removing the most decisive features first produces an early, steep decline, so a smaller area under the deletion curve (deletion AUC, given in the legend) indicates a more faithful explanation. Among the graph-based modules GAT (0.885) is more faithful than GCN (0.953). The MLP produces no importance weights; its features are removed in random order and it serves only as a control, so its value (0.691) is not a faithfulness measure and is not comparable with the graph-based modules.

For the GAT model, the area under the deletion curve was 0.885, whereas for GCN it was 0.953. A lower Deletion AUC indicates a larger decrease in model confidence when the most important features are removed and therefore reflects a stronger correspondence between the generated attributions and the underlying decision-making process.

These results suggest that the attention coefficients learned by GAT provide a more faithful representation of feature contributions than the uniform aggregation mechanism employed by GCN. The observed differences between the models remained consistent both on a subset of 5000 images and on the full dataset.

For the MLP model, the feature-removal order was selected randomly because the architecture does not

generate explicit feature importance weights. Consequently, the deletion results for MLP were used only as a baseline control and were not interpreted as a measure of attribution quality.

An example of an explanation path with a GPI_w value of 0.76 for an individual image is shown in Fig. 5.

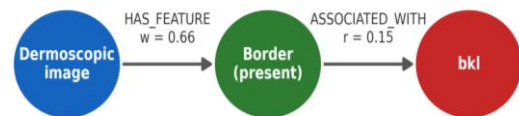


Fig. 5 Example of an explanation path for a single image: the path links the dermoscopic image to a present ABCD feature (edge HAS_FEATURE, weight w equal to the feature value) and then to the predicted diagnosis (edge ASSOCIATED_WITH, weight equal to the point-biserial coefficient r of the corresponding knowledge-graph edge). The path-level interpretability score for this decision is $GPI_w = 0.76$.

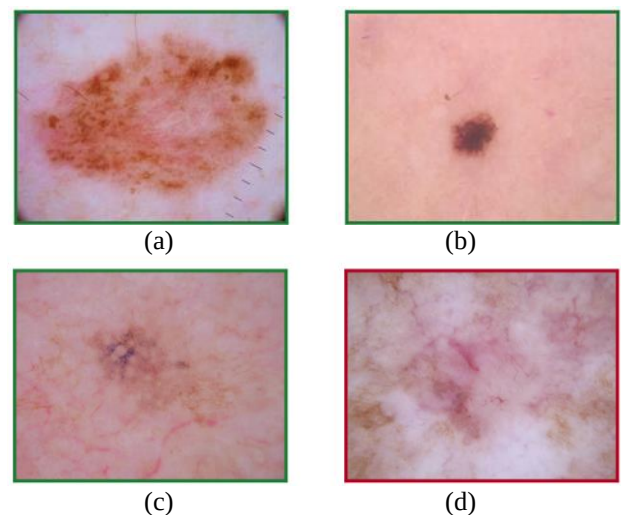


Fig. 6 Explanation paths of the GPI metric on four representative HAM10000 test images. Green frames indicate decisions with a valid explanation path based on a present ABCD feature significantly associated with the true diagnosis; the red frame indicates a case with no valid positive feature-diagnosis path in the ABCD knowledge graph. (a) melanoma explained by structural heterogeneity; (b) melanocytic nevus explained by color variation; (c) benign keratosis-like lesion explained by border irregularity; (d) basal cell carcinoma with no valid positive feature-diagnosis path. Feature abbreviations: A – asymmetry, B – border irregularity, C – color variation, D – diameter, S – structural heterogeneity; r is the point-biserial edge weight (Table 2).

To illustrate how the metric evaluates the ABCD rule on individual images, Fig. 6 shows four representative test cases. In the three lesions with a valid explanation path – melanoma via the Structure→melanoma edge ($r = 0.22$), melanocytic nevus via Color→nevus ($r = 0.14$) and benign keratosis via Border→bkl ($r = 0.15$) – the ABCD features present in the image

connect through a statistically significant edge to the correct diagnosis, so the decision is explainable and contributes to GPI_{cov} . For the basal cell carcinoma none of the present features has a significant positive edge to the bcc class (consistent with bcc showing only negative feature associations, Table 2), so no valid path exists and the metric correctly reports the decision as not explainable by the ABCD graph. These examples confirm that the metric operates directly on the ABCD-rule concepts and distinguishes explainable from non-explainable decisions.

3.5 External Validation on an Independent Dataset

To assess the generalization capability of the proposed approach, an additional evaluation was performed using the independent ISIC 2018 Task 3 dataset, which contains 1511 dermoscopic images from seven diagnostic categories. Model training was conducted on the HAM10000 dataset (8831 images for training and 1184 for validation, split by lesion_id), after which the model was evaluated on ISIC 2018 without further fine-tuning.

On the external dataset, the classifier achieved an Accuracy of 0.784, a Balanced Accuracy of 0.644, a macro-F1 score of 0.660, and a macro-AUC of 0.939. Compared with the results obtained on HAM10000 (0.807, 0.677, 0.668, and 0.955, respectively), only a moderate decrease in performance was observed, indicating that the model retained its effectiveness on previously unseen data.

Correlation analysis performed on the combined training data confirmed the main relationships between dermoscopic features and diagnostic categories. The strongest association remained that between structural heterogeneity and melanoma, with $r = 0.22$ and $q \approx 7 \times 10^{-93}$. In total, 13 statistically significant positive associations were identified.

The primary interpretability metric, GPI_{cov} , remained close to unity, indicating that explanation paths through the knowledge graph were preserved for the vast majority of predictions on the independent dataset as well (Fig. 7).

The conceptual reasoning modules based solely on five dermoscopic features exhibited a more pronounced decrease in performance on the independent dataset than the baseline classifier. This reduction is likely attributable to the limited information content of the compact feature set and differences in lesion segmentation quality. Whereas expert-provided lesion masks were available for HAM10000 images, segmentation masks for part of the external dataset were generated automatically using Otsu thresholding, which may have introduced additional errors into feature extraction.

Despite this reduction in predictive performance, the main findings related to interpretability remained unchanged. The knowledge graph reproduced the principal associations between dermoscopic features and diagnostic categories, and the value of GPI_{cov} remained close to unity. A more detailed investigation of the

robustness of GPI_w on independent datasets will require more reliable methods for extracting conceptual features and is considered a direction for future research.

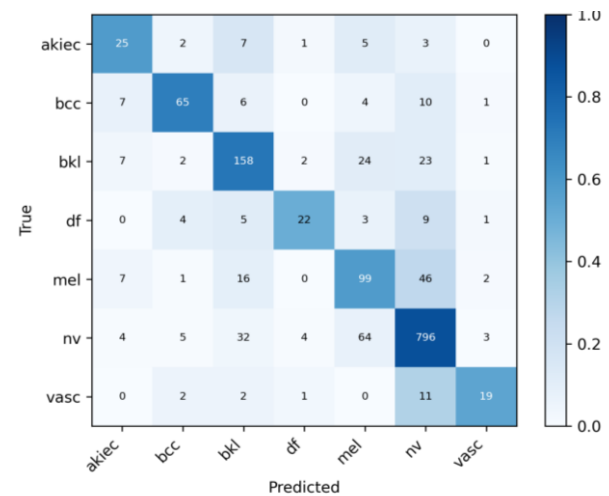


Fig. 7 Normalized confusion matrix for the classifier trained on HAM10000 and evaluated, without retraining, on the independent ISIC 2018 Task 3 test set ($n = 1511$). Rows are true and columns predicted diagnoses (row-normalized); the preserved diagonal structure indicates that the backbone generalizes across datasets. Class abbreviations as in Fig. 1.

4 Discussion

The results demonstrate that the interpretability of dermoscopic image classifiers can be quantified using diagnostic features routinely employed in clinical practice. Unlike saliency maps and other visual explanation techniques, the proposed approach links model decisions to specific ABCD-rule features and statistically validated associations between features and diagnoses. As a result, explanations are expressed in terms familiar to clinicians and can be directly related to established diagnostic criteria.

The problem of making deep dermoscopic classifiers interpretable has motivated a growing body of work. Post-hoc attribution methods such as saliency maps [18], LIME [19], class activation mapping [20], integrated gradients [21] and layer-wise relevance propagation [22] highlight influential pixels, whereas concept-based approaches such as TCAV [23], concept bottleneck models [15], prototype networks [24] and automatic concept extraction [25] relate predictions to higher-level concepts. In dermatology, multimodal and taxonomy-driven frameworks such as ExAID [26] and explainable diagnosis using taxonomies [27] generate concept-level justifications, and recent surveys of dermoscopic feature analysis [28] and of medical explainable AI [29] emphasise that interpretability should be expressed through the diagnostic criteria that clinicians already use. The present knowledge-graph formulation follows this line but adds a quantitative, model-agnostic score rather than a purely qualitative explanation.

The proposed metrics capture different aspects of interpretability. GPI_{cov} reflects the existence of at least one explanation path through the knowledge graph, whereas GPI_w quantifies the extent to which such paths contribute to the final prediction. In the present study, explanation paths were identified for nearly all predictions, both on the original dataset and during external validation. Differences between models were observed primarily in the degree to which these paths contributed to decision making.

The proposed framework distinguishes between models that explicitly exploit conceptual features through a knowledge graph and models that operate without a graph-based representation of knowledge. Under comparable classification performance, graph-based models exhibited a stronger correspondence between model decisions and clinically meaningful explanation paths. The sequential feature-removal analysis further confirmed the consistency of these explanations with model behaviour. Here the graph-based modules used graph attention [16] and graph convolution [30] to aggregate the conceptual features.

The faithfulness of explanations was assessed by sequential concept removal, following the rationale of remove-and-evaluate protocols [17, 31] and their known caveats [32]. Such perturbation-based measures, together with rationale-alignment benchmarks [33] and quantitative evaluations of saliency [34], have become standard tools for verifying that a highlighted feature is actually used by the model. In our setting the deletion analysis separated the graph-based reasoning modules from the black-box baseline, indicating that the GPI paths reflect genuine model behaviour rather than a post-hoc rationalisation.

The use of conceptual features did not lead to a substantial reduction in classification performance compared with models lacking an explicit knowledge representation. This observation is consistent with previous studies on concept-based learning and suggests that interpretability and predictive performance can be combined within a single architecture.

The proposed method complements existing explainable artificial intelligence approaches. Techniques such as Grad-CAM identify image regions that influence model predictions, whereas the graph-based approach relates predictions to diagnostic features and their relationships with diagnostic categories. Combining these complementary forms of explanation represents a promising direction for future research.

From the perspective of clinical deployment, the practical value of the proposed approach lies in expressing each decision through the ABCD criteria [12, 35] that dermatologists apply routinely, which may lower the barrier to trust and adoption relative to opaque predictions [36]. Modern computer-aided diagnosis systems, built on deep convolutional [37, 38] and transformer [39] architectures, already match or exceed dermatologist performance on curated benchmarks [2, 3, 40, 41], and classifiers trained on large clinical image collections have reached specialist-level accuracy

[42, 43]. Importantly, the largest gains in prospective studies arise from human–computer collaboration rather than from autonomous prediction [44]; an interpretability score that exposes whether a model reasons along clinically validated feature–diagnosis paths is directly useful in this collaborative setting. Optical modalities beyond dermoscopy – hyperspectral imaging [5, 6], combined dermoscopic-spectral analysis [7] and Raman or autofluorescence spectroscopy [8] – are increasingly used for skin-cancer diagnosis, and the modality-independent GPI metric could complement them by quantifying how far a model’s reasoning follows established feature–diagnosis relationships. Rather than competing with existing diagnostic pipelines, the metric is intended as an auxiliary, quantitative interpretability criterion reported alongside accuracy when a model is selected or certified for clinical use.

Several limitations should be acknowledged. The study relied on a simplified set of conceptual features derived from the ABCD rule, feature quality depended on lesion-segmentation accuracy, and evaluation was performed on a limited number of datasets. Expanding the feature vocabulary to include more detailed dermoscopic structures and conducting multi-centre validation studies may further improve the practical value of the approach. Future work will also investigate the applicability of the proposed metric to other medical imaging tasks and compare it with existing interpretability assessment methods. Broader validation on additional public collections such as the ISIC 2018 challenge data [45], BCN20000 [46] and seven-point-checklist datasets [47], together with formal notions of interpretability [48], is a natural next step.

5 Conclusions

A GPI metric was proposed for the quantitative assessment of interpretability in dermoscopic image classification models. The metric is based on the analysis of paths linking an image, dermoscopic features, and diagnostic categories through a knowledge graph.

Experiments on the HAM10000 dataset (10015 images) showed that the EfficientNet-B0 classifier reached an accuracy of 0.807, a balanced accuracy of 0.677, a macro-F1 of 0.668 and a macro-AUC of 0.955. A knowledge graph built from ABCD-rule features and comprising 11 statistically significant positive feature–diagnosis associations ($q < 0.05$) provided explanation paths for virtually all classifier decisions ($GPI_{cov} = 1.0$). Under comparable classification accuracy (0.67–0.68 for all reasoning modules), the proposed metric clearly separated models that explicitly use conceptual features ($GPI_w = 0.416$ for GAT and 0.559 for GCN) from the graph-free baseline ($GPI_w = 0$ for MLP). The sequential concept-removal analysis confirmed the faithfulness of the explanations, the deletion AUC being 0.885 for GAT and 0.953 for GCN versus 0.691 for MLP.

External validation on the independent ISIC 2018 dataset reproduced the principal feature–diagnosis associations and preserved interpretability

($GPI_{cov} \approx 1.0$), while classification accuracy decreased to 0.784. These results demonstrate the feasibility of quantitatively assessing interpretability using diagnostic features employed in clinical practice. The proposed GPI metric can serve as an additional criterion for the development and comparison of artificial intelligence models for dermoscopic image analysis.

Disclosures

All authors declare that there is no conflict of interest in this paper.

References

1. G. Argenziano, H. P. Soyer, “[Dermoscopy of pigmented skin lesions – a valuable tool for early](#),” *The Lancet Oncology* 2(7), 443–449 (2001).
2. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “[Dermatologist-level classification of skin cancer with deep neural networks](#),” *Nature* 542(7639), 115–118 (2017).
3. T. J. Brinker, A. Hekler, A. H. Enk, 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).
4. P. Tschandl, C. Rosendahl, and H. Kittler, “[The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions](#),” *Scientific Data* 5(1), 180161 (2018).
5. B. V. Grechkin, V. O. Vinokurov, Yu. A. Khristoforova, and I. A. Matveeva, “[VGG convolutional neural network classification of hyperspectral images of skin neoplasms](#),” *Journal of Biomedical Photonics & Engineering* 9(4), 040304 (2023).
6. S. N., V. A. R., S. R., K. N., P. M. M., and J. K., “[Highly Accurate Skin Cancer Diagnosis Using HMT-NET and Vision Transformer Models](#),” *Journal of Biomedical Photonics & Engineering* 11(4), 040308 (2025).
7. I. A. Matveeva, A. I. Komlev, O. I. Kaganov, A. A. Moryatov, and V. P. Zakharov, “[Multidimensional analysis of dermoscopic images and spectral information for the diagnosis of skin tumors](#),” *Journal of Biomedical Photonics & Engineering* 10(1), 010307 (2024).
8. I. A. Bratchenko, Yu. A. Khristoforova, L. A. Bratchenko, A. A. Moryatov, S. V. Kozlov, E. G. Borisova, and V. P. Zakharov, “[Optical biopsy of amelanotic melanoma with Raman and autofluorescence spectra stimulated by 785 nm laser excitation](#),” *Journal of Biomedical Photonics & Engineering* 7(2), 020308 (2021).
9. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bannetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “[Explainable Artificial Intelligence \(XAI\): Concepts, taxonomies, opportunities and challenges toward responsible AI](#),” *Information Fusion* 58, 82–115 (2020).
10. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “[Grad-CAM: visual explanations from deep networks via gradient-based localization](#),” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, 618–626 (2017).
11. S. M. Lundberg, S.-I. Lee, “[A unified approach to interpreting model predictions](#),” in *Advances in Neural Information Processing Systems* 30, 4765–4774 (2017).
12. F. Nachbar, W. Stolz, T. Merkle, A. B. Coggnetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig, “[The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions](#),” *Journal of the American Academy of Dermatology* 30(4), 551–559 (1994).
13. A. N. Zhdanova, “[Ontological approach to predicting psychological characteristics of social media users using knowledge graphs](#),” *Artificial Intelligence and Decision Making* 3, 40–59 (2025). [in Russian]
14. M. Tan, Q. V. Le, “[EfficientNet: rethinking model scaling for convolutional neural networks](#),” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR 97, 6105–6114 (2019).
15. P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “[Concept bottleneck models](#),” in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, *Proceedings of Machine Learning Research* 119, 5338–5348 (2020).
16. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “[Graph Attention Networks](#),” *arXiv preprint arXiv:1710.10903* (2017).
17. V. Petsiuk, A. Das, and K. Saenko, “[RISE: Randomized Input Sampling for Explanation of Black-box Models](#),” in *Proceedings of the British Machine Vision Conference (BMVC)* (2018).
18. K. Simonyan, A. Vedaldi, and A. Zisserman, “[Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps](#),” *arXiv preprint arXiv:1312.6034* (2013).
19. M. T. Ribeiro, S. Singh, and C. Guestrin, “[“Why Should I Trust You?”: Explaining the Predictions of Any Classifier](#),” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, 2016)*, pp. 1135–1144.
20. B. Zhou, A. Khosla, A. Lapedrizza, A. Oliva, and A. Torralba, “[Learning deep features for discriminative localization](#),” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929 (2016).

21. M. Sundararajan, A. Taly, and Q. Yan, “[Axiomatic attribution for deep networks](#),” in Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70, 3319–3328 (2017).
22. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “[On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation](#),” PLoS ONE 10(7), e0130140 (2015).
23. B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viégas, and R. Sayres, “[Interpretability beyond feature attribution: quantitative testing with concept activation vectors \(TCAV\)](#),” in Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80, 2668–2677 (2018).
24. C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. Su, “[This looks like that: deep learning for interpretable image recognition](#),” in Advances in Neural Information Processing Systems 32, 8930–8941 (2019).
25. A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, “[Towards automatic concept-based explanations](#),” in Advances in Neural Information Processing Systems 32, 9273–9282 (2019).
26. A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed, “[ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions](#),” Computer Methods and Programs in Biomedicine 215, 106620 (2022).
27. C. Barata, M. E. Celebi, and J. S. Marques, “[Explainable skin lesion diagnosis using taxonomies](#),” Pattern Recognition 110, 107413 (2021).
28. C. Barata, M. E. Celebi, and J. S. Marques, “[A survey of feature extraction in dermoscopy image analysis of skin cancer](#),” IEEE Journal of Biomedical and Health Informatics 23(3), 1096–1109 (2019).
29. E. Tjoa, C. Guan, “[A survey on explainable artificial intelligence \(XAI\): toward medical XAI](#),” IEEE Transactions on Neural Networks and Learning Systems 32(11), 4793–4813 (2021).
30. T. N. Kipf, M. Welling, “[Semi-Supervised Classification with Graph Convolutional Networks](#),” arXiv preprint arXiv:1609.02907 (2016).
31. S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, “[A benchmark for interpretability methods in deep neural networks](#),” in Advances in Neural Information Processing Systems 32, 9737–9748 (2019).
32. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “[Sanity checks for saliency maps](#),” in Advances in Neural Information Processing Systems 31, 9505–9515 (2018).
33. J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, “[ERASER: a benchmark to evaluate rationalized NLP models](#),” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics, 2020), pp. 4443–4458.
34. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, “[Evaluating the Visualization of What a Deep Neural Network Has Learned](#),” IEEE transactions on neural networks and learning systems 28(11), 2660–2673 (2017).
35. G. Argenziano, H. P. Soyer, S. Chimenti, et al., “[Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet](#),” Journal of the American Academy of Dermatology 48(5), 679–693 (2003).
36. C. Rudin, “[Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead](#),” Nature Machine Intelligence 1(5), 206–215 (2019).
37. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “[ImageNet classification with deep convolutional neural networks](#),” in Advances in Neural Information Processing Systems 25, 1097–1105 (2012).
38. K. He, X. Zhang, S. Ren, and J. Sun, “[Deep residual learning for image recognition](#),” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016).
39. A. Dosovitskiy, L. Beyer, A. Kolesnikov, “[An image is worth 16×16 words: transformers for image recognition at scale](#),” arXiv preprint arXiv:2010.11929 (2021).
40. H. A. Haenssle, C. Fink, R. Schneiderbauer, et al., “[Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists](#),” Annals of Oncology 29(8), 1836–1842 (2018).
41. P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, A. Lallas, J. Lapins, C. Longo, J. Malvehy, M. A. Marchetti, A. Marghoob, S. Menzies, A. Oakley, J. Paoli, S. Puig, C. Rinner, C. Rosendahl, A. Scope, C. Sinz, H. P. Soyer, L. Thomas, I. Zalaudek, and H. Kittler, “[Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study](#),” The Lancet Oncology 20(7), 938–947 (2019).
42. S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, “[Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm](#),” Journal of Investigative Dermatology 138(7), 1529–1538 (2018).
43. L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, “[Automated melanoma recognition in dermoscopy images via very deep residual networks](#),” IEEE Transactions on Medical Imaging 36(4), 994–1004 (2017).
44. P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, and H. Kittler, “[Human-computer collaboration for skin cancer recognition](#),” Nature Medicine 26(8), 1229–1234 (2020).

45. N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, “[Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration \(ISIC\)](#),” arXiv preprint arXiv:1902.03368 (2019).
46. M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, “[BCN20000: Dermoscopic Lesions in the Wild](#),” arXiv preprint arXiv:1908.02288 (2019).
47. J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, “[Seven-point checklist and skin lesion classification using multitask multimodal neural nets](#),” IEEE Journal of Biomedical and Health Informatics 23(2), 538–546 (2019).
48. F. Doshi-Velez, B. Kim, “[Towards A Rigorous Science of Interpretable Machine Learning](#),” arXiv preprint arXiv:1702.08608 (2017).