Journal of Healthcare Informatics Research (2023) 7:59–83
https://doi.org/10.1007/s41666-023-00127-4

RESEARCH ARTICLE

Interpretable Skin Cancer Classification based on Incremental Domain Knowledge Learning

Eman Rezk[1] · **Mohamed Eltorki**[2] · **Wael El-Dakhakhni**[1]

Received: 21 July 2022 / Revised: 2 January 2023 / Accepted: 3 February 2023 /
Published online: 15 February 2023

Abstract

The recent advances in artificial intelligence have led to the rapid development of computer-aided skin cancer diagnosis applications that perform on par with dermatologists. However, the black-box nature of such applications makes it difficult for physicians to trust the predicted decisions, subsequently preventing the proliferation of such applications in the clinical workflow. In this work, we aim to address this challenge by developing an interpretable skin cancer diagnosis approach using clinical images. Accordingly, a skin cancer diagnosis model consolidated with two interpretability methods is developed. The first interpretability method integrates skin cancer diagnosis domain knowledge, characterized by a skin lesion taxonomy, into model development, whereas the other method focuses on visualizing the decision-making process by highlighting the dominant of interest regions of skin lesion images. The proposed model is trained and validated on clinical images since the latter are easily obtainable by non-specialist healthcare providers. The results demonstrate the effectiveness of incorporating lesion taxonomy in improving model classification accuracy, where our model can predict the skin lesion origin as melanocytic or non-melanocytic with an accuracy of 87%, predict lesion malignancy with 77% accuracy, and provide disease diagnosis with an accuracy of 71%. In addition, the implemented interpretability methods assist understand the model's decision-making process and detecting misdiagnoses. This work is a step toward achieving interpretability in skin cancer diagnosis using clinical images. The developed approach can assist general practitioners to make an early diagnosis, thus reducing the redundant referrals that expert dermatologists receive for further investigations.

Keywords Artificial intelligence · Clinical images · Domain knowledge · Interpretability · Skin cancer · Skin lesion taxonomy

✉ Eman Rezk
rezke@mcmaster.ca

[1] School of Computational Science and Engineering, McMaster University, Hamilton, ON, Canada

[2] Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

1 Introduction

Skin cancer is one of the most common types of cancer in the USA; in 2022, it is estimated that 99,780 new invasive melanoma and 97,920 in situ melanoma cases have been diagnosed and 7650 cases died from the disease [1, 2]. However, early diagnosis dramatically improves the five-year survival rate, for example, in melanoma the 5-year survival rate for early diagnosed cases is 99% compared to 68% when melanoma reaches a nearby lymph node and 30% when melanoma spreads to distant lymph nodes [3].

Factors leading to delays or inaccurate diagnosis include a shortage of dermatologists [4] especially in rural areas [5], a gap that is subsequently filled by non-specialists, such as primary care providers who are inadequately trained to deal with complex and ambiguous dermatological conditions [6, 7] without access to diagnostic aids. Given the increasing prevalence of skin cancer and the chronic lack of dermatological expertise, there is a critical need to develop computer-aided skin cancer diagnostic applications that offer an accurate rapid early diagnosis.

The increased computational power and recent advances in artificial intelligence (AI) methods, such as deep learning, empowered the development of deep neural networks that can perform skin cancer diagnosis comparable and even superior to that of dermatologists [8–11]. Although deep learning applications have witnessed rapid expansion, the black-box nature and lack of robustness of such applications hinder their proliferation in clinical settings as they do not provide explanations of their decision-making process and any perturbations to the input can dramatically impact their performance and completely change the output [12–14].

In 2021, the European Commission proposed a set of rules to regulate the deployment of AI applications in the European market. Transparency was one of the requirements that AI applications, in particular high-risk applications such as medical services, should comply with [15]. As a consequence, the AI community created the explainable artificial intelligence (XAI) concept that aims at developing interpretability approaches to improve the transparency of the decision-making process, hence enhancing relevant human understandability and trust [16][16].

Numerous approaches have been investigated to enhance the interpretability of deep learning applications [18–21], where interpretability can be incorporated into the diagnostic model training or after training. In dermatology, deep neural networks (DNN) are trained to be interpretable by mimicking dermatologists' diagnosis strategy, as an example, a DNN was trained to perform hierarchical diagnosis and thus predict the lesion origin or lesion malignancy prior to making the disease diagnosis [22]. On the other hand, after training interpretability aims to integrate visualization tools with the developed model to rationalize its diagnosis. For example, in a skin lesion segmentation and classification task [23], a visualization map was integrated with the model to illustrate the image regions that were mostly used by the model to make a prediction. This would give the end-user confidence to know that the AI model was in fact analyzing the pathology on the provided image.

Concerning image modality, the number of *dermoscopic* image-based applications is substantially greater than those utilizing *clinical* images. In a recent review of XAI articles on skin cancer diagnosis [13], out of the 37 studies included in the review, 27 articles analyzed dermoscopic images, four articles utilized clinical images, two articles used both modalities, and four articles employed histopathological images. Dermoscopic images are skin images captured by a specialist using a special instrument (dermatoscope) to provide a magnified view of the lesion [24]. However, clinical images are skin images captured by a digital camera, thus obtaining clinical images do not require specific experience and acquiring high-quality clinical images is facilitated by the rapidly evolving smartphone cameras [25]. Given the fact that the burden of early diagnosis in dermatology relies on general practitioners [6] who can rarely perform dermoscopy [26], clinical images, rather than dermoscopic, are more relevant for skin lesion screening and thus the need for related XAI research.

In this work, we develop an interpretable skin cancer early diagnosis approach using *clinical* images. The proposed approach aims to (1) provide a low-cost rapid screening that can accelerate skin cancer diagnosis and (2) assist general practitioners in providing diagnosis prompting early referrals to dermatologists. Our proposed approach first incorporates domain knowledge, through a skin lesion taxonomy, into the design of a DNN to incrementally learn dermatological concepts. In addition, visual explanations of the diagnosis-making process are offered through advanced visualization maps to understand the rationale behind the model's final diagnosis. More specifically, the contributions of this work are:

1. Develop an incremental multi-output model that predicts the lesion origin (as melanocytic or non-melanocytic), classifies the lesion malignancy (malignant or benign), and provides a disease diagnosis (melanoma, nevi, basal cell carcinoma, and seborrheic keratosis).
2. Integrate two interpretability approaches to improve the transparency of the proposed model.
3. Investigate different loss functions for training the incremental model.
4. Implement several data balancing techniques.
5. Evaluate the proposed model and compare it with similar models.

2 Background

In this section, we explain some of the recent works that have tackled the problem of interpretability in skin cancer diagnosis. The work discussed herein is grouped by image modality into dermoscopic and clinical image interpretability approaches.

2.1 Dermoscopic Image Interpretability

Image similarity is an interpretability approach that simulates a dermatologist's diagnosis of new cases based on the knowledge of similar past cases. Barata et al. [27]

utilized a content-based image retrieval component when training a DNN model to diagnose melanoma. As such, the model was able to provide a diagnosis and retrieve similar images that justify the predicted diagnosis. Similarly, Codella et al. [28] utilized a hierarchy to group similar images when training a DNN to diagnose melanoma; consequently, the model was able to diagnose melanoma and retrieve similar images based on the similarity hierarchy. As a result, the model's prediction was accompanied by a set of similar images to justify the attained diagnosis.

The *ABCD* rule of dermoscopy is a well-established rule where asymmetry, border irregularity, color variations, and diameter features are analyzed to detect melanoma [29]. Chowdhury et al. [30] developed a melanoma diagnosis machine learning model that employs the *ABCD* features extracted using image transformations. In addition, a DNN consolidated with visual attention components was trained to diagnose melanoma and generate a visualization of the model's diagnosis process. The visualization of the DNN was found to be correlated with the output of the *ABCD* feature-based model indicating that the DNN implicitly learned the *ABCD* features and thus the DNN results can be trusted. Likewise, Stieler et al. [31] embedded the *ABCD* rule in the local interpretable model-agnostic explanations (LIME) [32] to develop an explainable melanoma diagnosis DNN model. LIME, a local surrogate model, was adapted such that its logic was replaced by the *ABCD* rule and trained on the predictions of the black-box model to visually explain the rationale behind the prediction.

Concept activation vector (CAV) is an interpretability method that evaluates the correlation between human-defined concepts and the model's prediction. CAV is used as a deep learning model testing strategy that measures the importance of human-defined concepts in the results of the model [33]. Lucieri et al. [34] employed CAV to interpret the results of a deep learning model trained to diagnose melanoma, nevi, and seborrheic keratosis. They utilized several skin lesion features, such as lesion pigmentation, streaks, dots and globules, and blue-whitish veils, to represent the concepts for understanding the model's diagnosis. Their work showed that there is a strong correlation between the model's prediction and the explored skin lesion features which indicates that the model learned human understandable concepts.

Utilizing a skin lesion taxonomy is an interpretability approach that simulates the dermatologists in dividing the diagnosis task into a hierarchy of subtasks [35] based on various criteria such as lesion origin or malignancy before reaching the diagnosis. In 2019, Barata et al. [22] employed a two-level skin lesion hierarchy to develop two DNNs to classify an image as melanoma versus nevi and seborrheic keratosis versus nevi. Both networks were trained to learn one level of the hierarchy and then perform binary classification on the disease level without considering the dependency between the hierarchy levels. In 2021, Barata et al. [36] extended the taxonomy presented in [22] to include a three-level hierarchy implemented with a recurrent neural network trained on the features of dermoscopic images extracted by an encoder network.

Attribution maps, also known as saliency maps, are visualizations that outline the areas contributing to a diagnosis decision accordingly providing a visual explanation of the diagnosis-making process. The class activation map (CAM) [37] and

gradient-weighted class activation maps (Grad-CAM) [38] are popular techniques that were widely employed as DNN visual explanation tools [23, 39–43].

2.2 Clinical Image Interpretability

Compared to dermoscopic images, clinical images have been exhibiting much less attention in addressing the gap related to their interpretability; moreover, the employed XAI approaches mainly focused on providing visual justifications of the DNN output using variants of attribution map techniques. In 2017, Ge et al. [44] adapted the classical CAM to work on the bilinear pooling feature map [45] to provide detailed visual explanations of the diagnosis process. Later in 2018, Grad-CAM was implemented to improve the understandability of a DNN diagnosis of 12 skin lesions [46]. Pfau et al. [47] implemented an aggregated global visualization approach based on Grad-CAM and the competitive gradient input method [48] to study how the DNN model adapts to skin image artifacts, such as ink.

Furthermore, CAM was used to visualize the regions of the image that contribute to the predicted labels of a DNN developed by Kawahara et al. [49]. The visualizations were utilized to interpret the prediction of the 7-point criteria associated with melanoma [50] and the diagnosis of five skin lesions. Another attribution map technique, the integrated gradients [51], was deployed in [52] where the DNN was trained to mainly diagnose 26 skin conditions. Using integrated gradients demonstrated the model's learnability by highlighting the significant image pixels that led to the prediction. Additionally, the integrated gradients technique [51] was combined with SmoothGrad [53] to offer an averaged visualization map, over a set of images, that was then utilized to study the correlation between the model's learnt features and human-labeled region of interest in classifying skin lesions [54].

After revising the recent interpretability work in skin cancer diagnosis, it can be observed that the work implemented in clinical image interpretability is primarily based on providing visual explanations of the predictions. Therefore, we aim to develop an interpretable skin cancer diagnosis model, trained and validated on clinical images, that incorporates domain knowledge into model training and provides visual explanations of the predictions after training. Accordingly, we embed dermatology knowledge in our model training by implementing a well-established skin lesion taxonomy that mimics dermatologists in diagnosing the lesions based on the lesion's origin and malignancy. In addition, we utilize an attribution map technique to provide visual explanations of the predictions as an after-training interpretability approach.

It is worth mentioning that the taxonomy, as a source of dermatology knowledge, has been partially utilized in [22] where the authors used only two levels of the taxonomy and developed two separate models each providing a binary classification of the diseases (explained in Sec. 2.1). In addition, the taxonomy has been utilized in [36] where the full taxonomy was implemented using image encoding and a recurrent neural network. In our work, we implement the full taxonomy following a different approach, and we develop a single multi-output model that incrementally predicts the lesion origin, lesion malignancy, and the disease. In addition, we build

the taxonomy using a convolutional neural network with a customized loss function to help the model learn the dependency between the taxonomy levels and hence improve the disease diagnosis.

3 Data and Taxonomy

The clinical images of a publicly available dataset containing 1011 skin lesion cases [55] were employed in our work. Originally, the dataset was used to predict the 7-point criteria linked to melanoma and perform lesion diagnosis for 5 disease classes, basal cell carcinoma (BCC), nevi (NEV), melanoma (MEL), seborrheic keratosis (SK), and a miscellaneous class (MISC) that includes any other disease such as dermatofibroma [49] (the number of images for each disease is listed in Table 1). In our work, we focused on four classes, BCC, NEV, MEL, and SK to fit with the three-level skin lesion taxonomy [36] shown in Fig. 1.

The first level of the taxonomy identifies the lesion origin as melanocytic or non-melanocytic, the second level groups lesions based on malignancy as malignant or benign, and the final level leads to the disease. The total number of images included in our work is 914 images, distributed as presented in Fig. 2 across the taxonomy levels.

Table 1 Data description

Disease	No. clinical images	Class	No. clinical images
Basal cell carcinoma	42	BCC	42
Blue nevus	28	NEV	575
Clark nevus	399		
Combined nevus	13		
Congenital nevus	17		
Dermal nevus	33		
Recurrent nevus	6		
Reed or Spitz nevus	79		
Melanoma	1	MEL	252
Melanoma in situ	64		
Melanoma-less than 0.76 mm	102		
Melanoma between 0.76 and 1.5 mm	53		
Melanoma-greater than 1.5 mm	28		
Melanoma metastasis	4		
Seborrheic keratosis	45	SK	45
Dermatofibroma	20	MISC	97
Lentigo	24		
Melanosis	16		
Miscellaneous	8		
Vascular lesion	29		

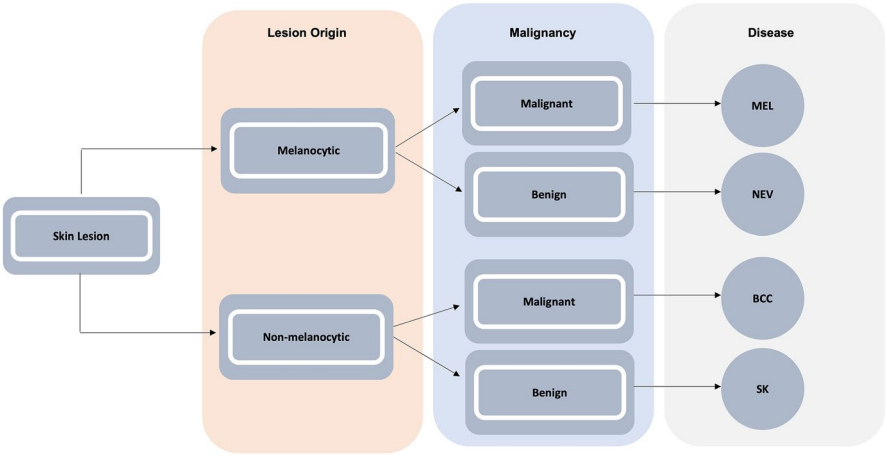

Fig. 1 Skin lesion taxonomy

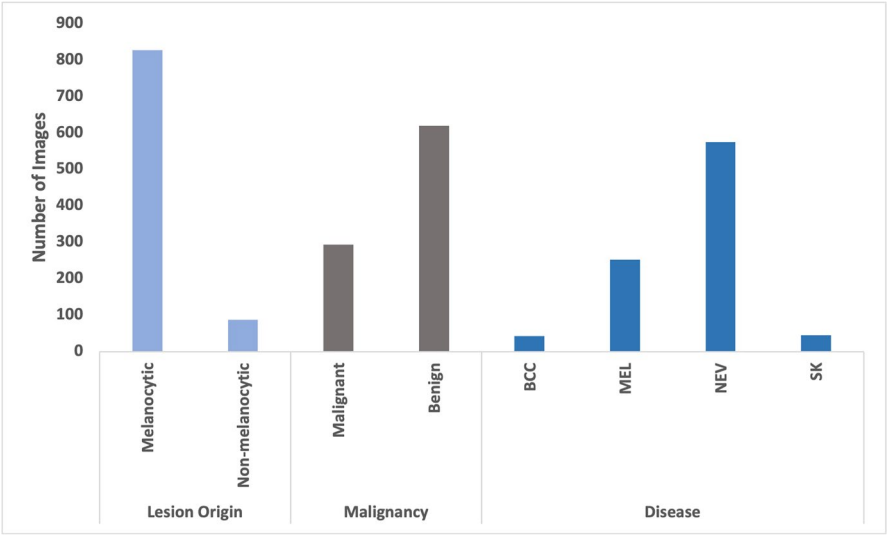

Fig. 2 Study data distribution

Based on Fig. 2, it can be observed that the data is imbalanced at all levels. At the lesion origin level, there are 827 melanocytic and 87 non-melanocytic; at the malignancy level, there are 294 malignant and 620 benign; and at the disease level, there are 42 BCC, 252 MEL, 575 NEV, and 45 SK cases. The imbalance ratio of the data is calculated as per Eq. 1 [56], the imbalance ratio for the levels 1, 2, and 3 is 9.51, 2.11, and 13.69, respectively.

$$Imbalance\ Ratio = \frac{N_{maj}}{N_{min}} \tag{1}$$

where N_{maj} and N_{min} are the numbers of majority and minority class instances, respectively.

4 Methods

In this work, we implemented two interpretability approaches to improve the transparency of skin cancer diagnosis. First, we adapted a DNN architecture to incrementally learn dermatological concepts and consequently increase the interpretability of the diagnosis-making process. Second, a visual saliency map is created to explain skin lesion predictions on the disease level. Since the utilized data is imbalanced as discussed in Sec. 3, we address this problem by developing several data balancing approaches.

4.1 Proposed Architecture

The Inception V3 architecture, introduced by Szegedy et al. [57], performed well in skin lesion diagnosis [49, 58]; we thus adapted the Inception V3 to mimic dermatologists in incrementally diagnosing a lesion by developing a multi-output incremental diagnosis network. The main blocks of the Inception V3 network (Fig. 3) were retained to extract features from skin images, whereas the classification layers of the network were replaced with the incremental diagnosis block to obtain domain knowledge represented as the skin lesion taxonomy.

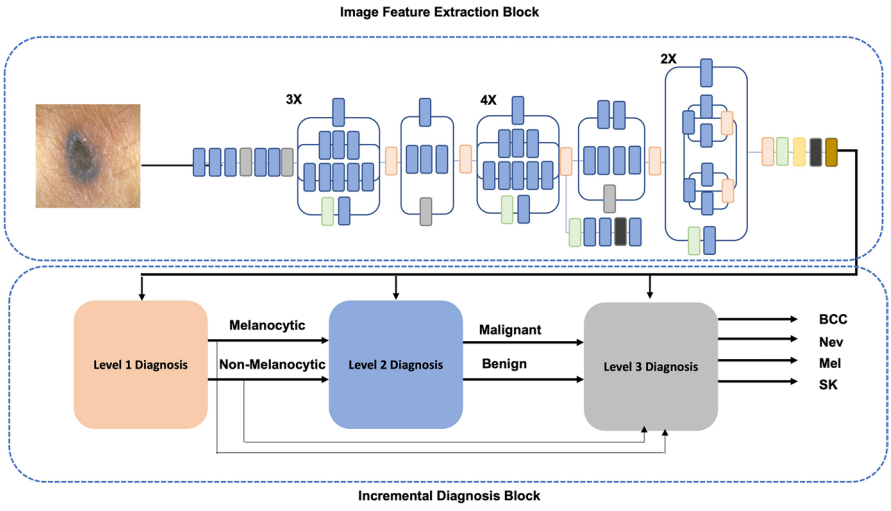

Fig. 3 Proposed network architecture

During training, the image features (output of the extraction block) were given as input to all diagnosis levels (Fig. 3). In addition, levels 2 and 3 were provided with the lesion origin and malignancy, as such the network was trained to predict the output of all levels. As a result, when making a diagnosis in the testing phase, the network could justify the final disease prediction based on level 1 and 2 predictions.

To optimize the diagnosis model during training, a loss function was used to calculate the prediction divergence from the actual diagnosis and update the model accordingly. The categorical cross entropy (CE) loss function (Eq. 2) is one of the most commonly used loss functions [59]. However, the CE considers the loss for each diagnosis level separately without considering the dependency between the levels. Subsequently, we adapted the CE loss to consider the skin lesion taxonomy levels and calculate the loss for each level while considering the previous level as shown in Eq. 3, the taxonomy CE loss (TCE).

$$Loss_{CE}(\hat{y}, y) = -\sum_{j=1}^{C} y_j \log \hat{y}_j \tag{2}$$

$$Loss_{TCE}(\hat{y}, y) = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij} \tag{3}$$

where y_{ij} is the actual diagnosis, $\hat{y}_{ij}$ is the predicted diagnosis, C is the number of classes in each level, and N is the levels.

4.2 Visual Saliency Maps

Visual saliency maps are visualizations that emphasize the pixels of an image that mostly influence the classification of a DNN. Gradient-based visualization methods are approaches that calculate the classification gradient given the input features [60] such as the widely used GradCAM which backpropagates the gradient to the last convolutional layer to create a fine-grained visualization map [60]. In our work, we utilized GradCAM++[61] which is an extension of the GradCAM where the former approach provides better visualization and more accurate object localization and and detects multiple occurrences of the objects [61].

4.3 Data Balancing

Based on the data distribution in Fig. 2 and the imbalance ratios calculated in Sect. 3, it can be inferred that there is a noticeable class imbalance in all taxonomy levels. As a result, the developed models are expected to be biased toward the majority classes [62]. Therefore, we implemented three techniques to address class imbalance on the algorithm and data levels. On the algorithm level, the class frequency-based weighted loss function [63] is used to handle class imbalance where each class is assigned a weight that is inversely proportional to the number of instances in that class. Therefore, the minority class receives higher weights compared to the

majority class, the class frequency-based weighted CE loss implemented herein for skin lesion diagnosis with the taxonomy is defined as Eq. 4.

$$Weighted\ Loss_{TCE}(\hat{y}, y) = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} W_{ij}\ y_{ij}\ \log\ \hat{y}_{ij} \tag{4}$$

where W_{ij} is the inverse of the number of instances in level i class j and the weights are normalized over the number of classes to balance the loss [64].

On the data level, we implemented two data augmentation approaches. First, data transformation, where image geometric transformations (i.e., flipping and rotation) were performed to increase the minority class size [65], and second, data integration in lieu of artificially expanding the minority class data size using transformations, we integrate real images from another image source, DermNet NZ [66], to augment the minority class.

5 Empirical Framework

In this section, we describe the empirical setup designed to investigate the proposed methods, then explain in detail model training configurations, and finally, clarify the evaluation metrics utilized herein to assess the developed models.

5.1 Empirical Setup

The empirical setup is designed to study the impact of the incremental architecture, the loss function, the data balancing, and the K-fold cross-validation on the disease diagnosis accuracy. As a result, our work incorporates four setups: (1) a baseline model that predicts the diseases directly to be compared with the incremental model that learns dermatological concepts based on a skin lesion taxonomy; (2) a model with the categorical CE loss to be compared with a model with the TCE loss; (3) three models with class frequency-based weighted loss function, data transformation, and data integration are developed for investigating data balancing; and (4) three models implementing K-fold cross-validation [67] ($K=5$) for class frequency-based weighted loss function, data transformation, and data integration. In all setups, the significance of the difference in disease level accuracy across the developed models is measured using the p value of the Z-test for comparing two proportions with a 95% confidence level [68].

In the third setup, data transformation and integration are applied to the minority classes. In data transformation, training images belonging to the minority class diseases, SK and BCC, were replaced by three transformed images created by applying random flipping, rotation, and adding Gaussian noise. Consequently, the training data size increased to 652 images, and the validation and test data sizes did not change. In data integration, 42 SK images and 152 BCC images collected

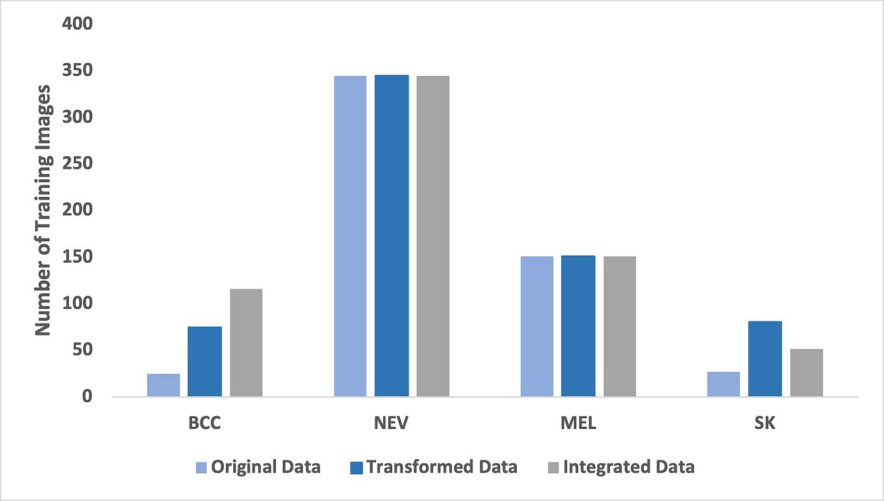

Fig. 4 Training data class distribution

from DermNet NZ were consolidated with the study data to augment the minority classes. Unlike data transformation, integrating data from different sources necessitates redistributing the images across training, validation, and testing. As a result, the total number of images (1108) was split into 60% training (664), 20% validation (222), and 20% testing (222 images). Figure 4 illustrates the disease distribution of the original, transformed, and integrated training dataset.

5.2 Model Training

The study data, 914 images, were mainly split into 60% (548) training, 20% (183) validation, and 20% (183) testing based on the split provided by Kawahara et al. [49] of the utilized 7-point criteria evaluation database [55] for proper performance comparison. However, in the k-fold cross-validation experimental setup, a fivefold data split was implemented on the training and validation (731 images), while model testing was performed on the independent test set (183 images). The baseline and the incremental architectures are built using the Inception V3 network [57] while applying transfer learning [69] to benefit from the network weights gained from training on ImageNet [70]. To refine the Inception V3 for the skin lesion diagnosis task, we replaced the classification layers of the network with a global average pooling layer and Softmax layers [71]. In the baseline model, only one Softmax layer with 4 output units was added to the network to directly perform disease diagnosis (MEL, NEV, BCC, SK).

In the incremental architecture, first, a Softmax layer with input as the image features and 2 output units representing the first level of the taxonomy (melanocytic or non-melanocytic) was added. Subsequently, we added a concatenate layer

to combine the image features with the classification output from the first Softmax layer where the output of the concatenation layer was fed as input to a second Softmax layer with 2 output units which is responsible to perform the classification of the second level of the taxonomy as malignant or benign. Finally, we added another concatenation layer to combine image features with the classification outputs from the first and second Softmax layers, accordingly, and a third Softmax layer with 4 output units was added to use that integrated input to perform disease level classification (MEL, NEV, BCC, SK).

Online image augmentation, such as image flipping, rotation, zoom, and shift, was performed during training. To calibrate the pre-trained Inception V3 with the skin lesion diagnosis task, we unfreeze the last two blocks of the network to allow them to train with the skin lesion classification layers for 50 epochs with batch size 32. In model optimization, we utilized the stochastic gradient descent with a learning rate of 0.001 and momentum of 0.9 for consistently reduced loss, fast convergence, and decreased oscillations [72]; all experiments were implemented using Keras [73] for model training and optimization, while TensorFlow [74] was used for class label transformation from categorical to numerical values.

5.3 Evaluation

All developed models were evaluated in terms of accuracy, sensitivity (recall), specificity, precision, and F-score metrics [75] as per Eqs. [5–9].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Sensetivity/Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$F - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{9}$$

In the equations above, *TP* is the true positives (count of correctly classified positive class), *TN* is the true negatives (count of correctly classified negative class), *FP* is the false positives (count of instances incorrectly classified as a positive class), and *FN* is the false negatives (count of instances incorrectly classified as a negative class). In level 1 of the taxonomy, the positive class is the melanocytic, and the negative class is the non-melanocytic. In level 2, the malignant class is the positive, and the benign is the negative class. In level 3, each disease represents a positive class, and all others are the negative class.

Table 2 Setup-1 and setup-2 models: the F-score is undefined (U) when sensitivity and precision are zero

Approach	Class	Sensitivity	Specificity	Precision	F-score	Overall accuracy
Baseline	BCC	0	1	0	U	0.58
	NEV	0.92	0.01	0.61	0.73	
	MEL	0.02	0.93	0.10	0.03	
	SK	0	1	0	U	
Incremental (M_1)	Lesion origin	1	0	0.91	0.95	0.91
	Malignancy	0.02	1	1	0.04	0.68
	BCC	0	1	0	U	0.39
	NEV	0.24	0.81	0.68	0.35	
	MEL	0.86	0.26	0.31	0.46	
	SK	0	1	0	U	
	Average					0.66
TCE (M_2)	Lesion origin	0.98	0	0.91	0.94	0.89
	Malignancy	0.58	0.81	0.59	0.58	0.73
	BCC	0	1	0	U	0.69
	NEV	0.89	0.47	0.74	0.81	
	MEL	0.47	0.86	0.56	0.51	
	SK	0	0.99	0	U	
	Average					0.77

6 Results

In the first empirical setup, summarized in Table 2, it can be noticed that the baseline model has an overall accuracy (0.58) higher than the disease level overall accuracy (0.39) in the incremental model (M_1). However, the sensitivity of the baseline model indicates that it has poor discrimination between the diseases; subsequently, all the test images are classified as NEV because it is the majority class. On the other hand, in the incremental model, we can see a slight improvement in the model discrimination, on the disease level, between MEL (sensitivity 0.86 and precision 0.31) and NEV (sensitivity 0.24 and precision 0.68) as a result of the knowledge gained from level 2 of the taxonomy. When observing the sensitivity and specificity of the baseline and M_1 (Table 2), we can find multiple occurrences of zero sensitivity and one specificity (or the opposite) indicating that both models suffer from the class imbalance that led to biased models toward the majority classes.

Turning to the second empirical setup where we study the impact of the taxonomy-based loss function (TCE) on the performance of the incremental architecture of M_1. Table 2 shows a comparison between the incremental model with categorical CE (M_1) and with TCE (M_2) loss functions. It can be inferred that the TCE significantly improved the overall model performance where the accuracy increased from 0.66 to 0.77. Moreover, there is a considerable enhancement in the classification accuracy in the malignancy prediction between the two models; thus, the accuracy increased from 0.68 to 0.73. Similarly, the accuracy of the disease prediction

improved from 0.39 to 0.69. To measure the significance of the change in disease classification accuracy across the three models, the p value of the Z test for comparing the accuracies with a 95% confidence level is reported. Comparing the baseline with M_1 and M_2, $p = 0.00038$ and 0.0226, respectively, for comparing M_1 and M_2, $p < 0.00001$. As such, all p values < 0.05 meaning that the differences in accuracies across the three models are significant.

As the model architecture does not impact the first level of the taxonomy, there is no noticeable improvement in the lesion origin performance and both models with CE and TCE are biased toward the majority class (melanocytic positive class), thus achieving a sensitivity of 1.00 (M_1) and 0.98 (M_2), whereas the specificity is zero.

When observing the sensitivity and specificity of M_1 and M_2 at the malignancy prediction level, we can find a noticeable balance between the sensitivity (0.58) and specificity (0.81) in the model with TCE compared to the model with CE loss (sensitivity of 0.02 and specificity of 1). This indicates the increased ability of the model with TCE loss to differentiate between the malignant and benign cases although the classes are imbalanced. Similarly, on the disease level, the model with TCE loss shows more balance between NEV and MEL sensitivity and specificity compared to the CE loss model. Nevertheless, none of the models was able to diagnose BCC or SK.

In the third empirical setup, data balancing approaches were implemented based on the outperforming model in the second setup (M_2) utilizing the TCE as the loss

Table 3 Setup-3 models: the F-score is undefined (U) when sensitivity and precision are zero

Approach	Class	Sensitivity	Specificity	Precision	F-score	Overall accuracy
Weighted loss	Lesion origin	0.92	0.35	0.93	0.92	0.86
	Malignancy	0.47	0.75	0.47	0.47	0.66
	BCC	0	0.99	0	U	0.57
	NEV	0.86	0.22	0.65	0.74	
	MEL	0.04	0.92	0.17	0.06	
	SK	0.23	0.92	0.18	0.20	
	Average					0.70
Data transformation	Lesion origin	0.98	0	0.91	0.94	0.89
	Malignancy	0.36	0.90	0.62	0.46	0.72
	BCC	0.12	0.99	0.50	0.19	0.67
	NEV	0.93	0.29	0.69	0.79	
	MEL	0.27	0.92	0.58	0.37	
	SK	0	0.99	0	U	
	Average					0.76
Data integration (M_3)	Lesion origin	0.95	0.61	0.88	0.91	0.86
	Malignancy	0.71	0.79	0.70	0.70	0.76
	BCC	0.74	0.95	0.74	0.74	0.68
	NEV	0.83	0.70	0.75	0.79	
	MEL	0.47	0.87	0.51	0.49	
	SK	0.24	0.98	0.44	0.31	
	Average					0.77

function. Class frequency-based weighted loss, data transformation, and data integration balancing techniques (explained in Sec. 4.3 and designed as detailed in Sec. 5) were implemented. The performance of the models addressing data imbalance is summarized in Table 3. Data transformation and data integration techniques improved disease classification accuracy compared to the weighted loss. The difference in accuracy between the weighted loss and data transformation and the weighted loss and data integration is significant ($p=0.04036$ and 0.02034, respectively). However, the difference in accuracy between the data transformation and data integration is not significant ($p=0.86502$).

It can be noticed that data integration balancing has the ability to differentiate between the melanocytic (sensitivity 0.95) and non-melanocytic (specificity 0.61) classes in the lesion origin level; similarly, the model can differentiate between malignant (sensitivity 0.71) and benign (specificity 0.79) lesions, and on the disease level, the model can diagnose all diseases with higher precision compared to the other models.

Comparing the weighted loss with the data transformation model, the former considerably improved the class imbalance; therefore, the model started to correctly classify the minority classes. This can be inferred by the specificity of 0.35 in lesion origin prediction which reflects the ability of the model to predict the non-melanocytic compared to zero specificity in the data transformation model (Table 3) and all other models (Table 2).

On the disease level, when observing the sensitivity, the weighted loss model successfully diagnosed a few SK cases but failed to diagnose the BCC images. Similarly, the data transformation model predicted a few cases of BCC and failed to predict the SK. It is important to note that the data integration model correctly diagnosed 0.24 of the SK cases compared to zero cases for the data transformation model although the number of SK images utilized in training the latter model is 81 which is more than the number of the SK images considered for training the former model (52 images). This observation emphasizes the benefit of using real images to augment the training instead of artificially altering the images.

In the last experimental setup, we implemented a five-fold cross-validation strategy for all data balancing techniques. As such the training and validation data are combined and divided into five folds, four utilized for training and one for validation and shuffled iteratively. Accordingly, five models are developed for each data balancing technique, and the average performance along with the standard deviation is reported in Table 4.

Although data transformation and data integration improved the diseases classification accuracy compared to the weighted loss technique, the difference in accuracy between the weighted loss and the data transformation is not significant ($p=0.1031$), but the difference in accuracy between the weighted loss and the data integration is significant ($p=0.01046$). Finally, the difference in accuracy between data transformation and data integration is not significant ($p=0.38978$).

With respect to the change in accuracy between all data balancing techniques with and without performing cross-validation (Tables 3 and 4), in the weighted loss, the disease accuracy insignificantly increased from 0.57 to 0.59 ($p=0.67448$). In the data transformation, the disease accuracy did not change (0.67); finally, in the

Table 4 Setup-4 models: the five-fold cross-validation (CV) (score ± standard deviation)

Approach	Class	Sensitivity	Specificity	Precision	F-score	Overall accuracy
CV weighted loss	Lesion origin	0.93 ± 0.02	0.32 ± 0.11	0.93 ± 0.01	0.93 ± 0.01	0.87 ± 0.01
	Malignancy	0.22 ± 0.06	0.90 ± 0.05	0.54 ± 0.09	0.31 ± 0.05	0.68 ± 0.02
	BCC	0.08 ± 0.10	0.97 ± 0.03	0.07 ± 0.08	0.16 ± 0.07	0.59 ± 0.01
	NEV	0.89 ± 0.01	0.30 ± 0.04	0.68 ± 0.01	0.77 ± 0.0	
	MEL	0.12 ± 0.05	0.90 ± 0.03	0.30 ± 0.07	0.16 ± 0.06	
	SK	0.07 ± 0.09	0.96 ± 0.02	0.05 ± 0.06	0.14 ± 0.03	
	All Levels					0.72 ± 0.01
CV data transformation	Lesion origin	0.98 ± 0.01	0.04 ± 0.05	0.91 ± 0.01	0.94 ± 0.00	0.89 ± 0.01
	Malignancy	0.45 ± 0.03	0.83 ± 0.03	0.56 ± 0.05	0.50 ± 0.03	0.71 ± 0.02
	BCC	0.15 ± 0.09	0.99 ± 0.00	0.40 ± 0.20	0.27 ± 0.07	0.67 ± 0.02
	NEV	0.87 ± 0.04	0.45 ± 0.04	0.73 ± 0.01	0.79 ± 0.02	
	MEL	0.41 ± 0.04	0.85 ± 0.04	0.52 ± 0.08	0.45 ± 0.04	
	SK	0.02 ± 0.04	0.99 ± 0.01	0.05 ± 0.09	0.05 ± 0.07	
	All Levels					0.75 ± 0.01
CV data integration (M_4)	Lesion origin	0.95 ± 0.01	0.63 ± 0.05	0.88 ± 0.02	0.92 ± 0.01	0.87 ± 0.01
	Malignancy	0.69 ± 0.03	0.83 ± 0.04	0.74 ± 0.04	0.71 ± 0.02	0.77 ± 0.02
	BCC	0.73 ± 0.02	0.97 ± 0.01	0.82 ± 0.04	0.77 ± 0.01	0.71 ± 0.02
	NEV	0.85 ± 0.02	0.71 ± 0.05	0.76 ± 0.03	0.80 ± 0.01	
	MEL	0.50 ± 0.06	0.85 ± 0.02	0.51 ± 0.03	0.50 ± 0.04	
	SK	0.31 ± 0.03	0.98 ± 0.01	0.63 ± 0.09	0.42 ± 0.03	
	All Levels					0.78 ± 0.01

data integration, the disease accuracy insignificantly increased from 0.68 to 0.71 ($p = 0.47152$). It is worth mentioning that, developing a model based on the five-fold cross-validation approximately requires five times the computational time of developing the same model without cross-validation. Accordingly, the selection between the k-fold cross-validation and the train-validation split is a trade-off between accuracy and computational time.

To summarize the results of all empirical setups, we created a class-specific performance comparison based on the sensitivity and specificity of each model. In Fig. 5, M_1 represents the incremental architecture; M_2 is the incremental architecture and the taxonomy-based loss function; M_3 includes the incremental architecture, the taxonomy-based loss function, and minority class data integration; and M_4 represents the incremental architecture, the taxonomy-based loss function, data integration, and five-fold cross-validation.

In lesion origin classification, all models were similarly able to predict the melanocytic class; however, only M_3 and M_4 were able to predict the non-melanocytic class. Thus, neither the incremental architecture nor the taxonomy-based loss function impacted the performance of lesion origin classification which completely abides by the logic of the developed architecture.

Regarding malignancy classification, M_1 pertains to the incremental architecture only, resulting in a biased model toward the benign class; nevertheless, M_2, M_3, and

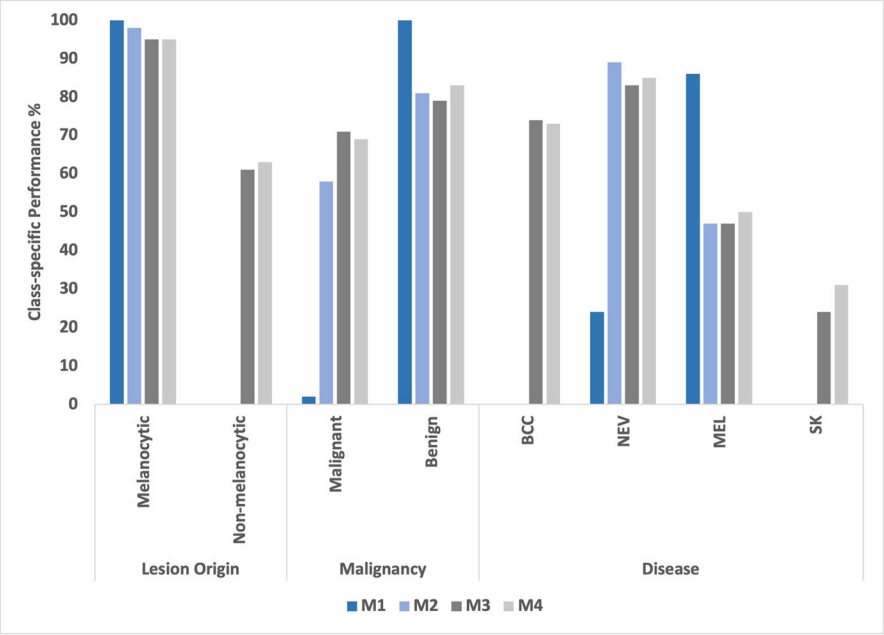

Fig. 5 Experiment summary

M_4 resulted in a noticeable balance in detecting malignant and benign cases. Finally, on the disease level, M_1 and M_2 were able to distinguish between MEL and NEV, but the underrepresented diseases BCC and SK were not captured by the models. Given the high imbalance ratio (13.69) at this taxonomy level, data integration for minority class balancing implemented in M_3 and M_4 substantially improved the diagnosis of BCC and SK.

Moving to the interpretability of the developed models, an example of the correctly diagnosed BCC, NEV, MEL, and SK cases is illustrated in Fig. 6a, b, c, and d. The lesion taxonomy clearly explains the rationale behind the final disease diagnosis. In addition, the presented probability of each disease group improves the transparency and confidence of the results.

The lesion taxonomy is also beneficial in detecting misdiagnoses as illustrated in Fig. 7 where a MEL lesion was misdiagnosed as NEV. This error can be discovered based on the malignancy prediction, as NEV is not a malignant lesion as shown by the model. Thus, the model's built-in interpretability facilitated understanding the output of the model. However, there are cases where the taxonomy will not help detect the misdiagnosis as shown in Fig. 8, a case of NEV misdiagnosed as MEL. The lesion is diagnosed correctly as melanocytic but misdiagnosed as malignant and MEL; since logically all the predicted taxonomy paths are correct, this misdiagnosis cannot be detected. This error is expected to happen in any computer-aided diagnosis application.

Finally, we present, in Fig. 9, the second interpretability approach of Grad-CAM++to highlight the influential regions that contributed most to the disease

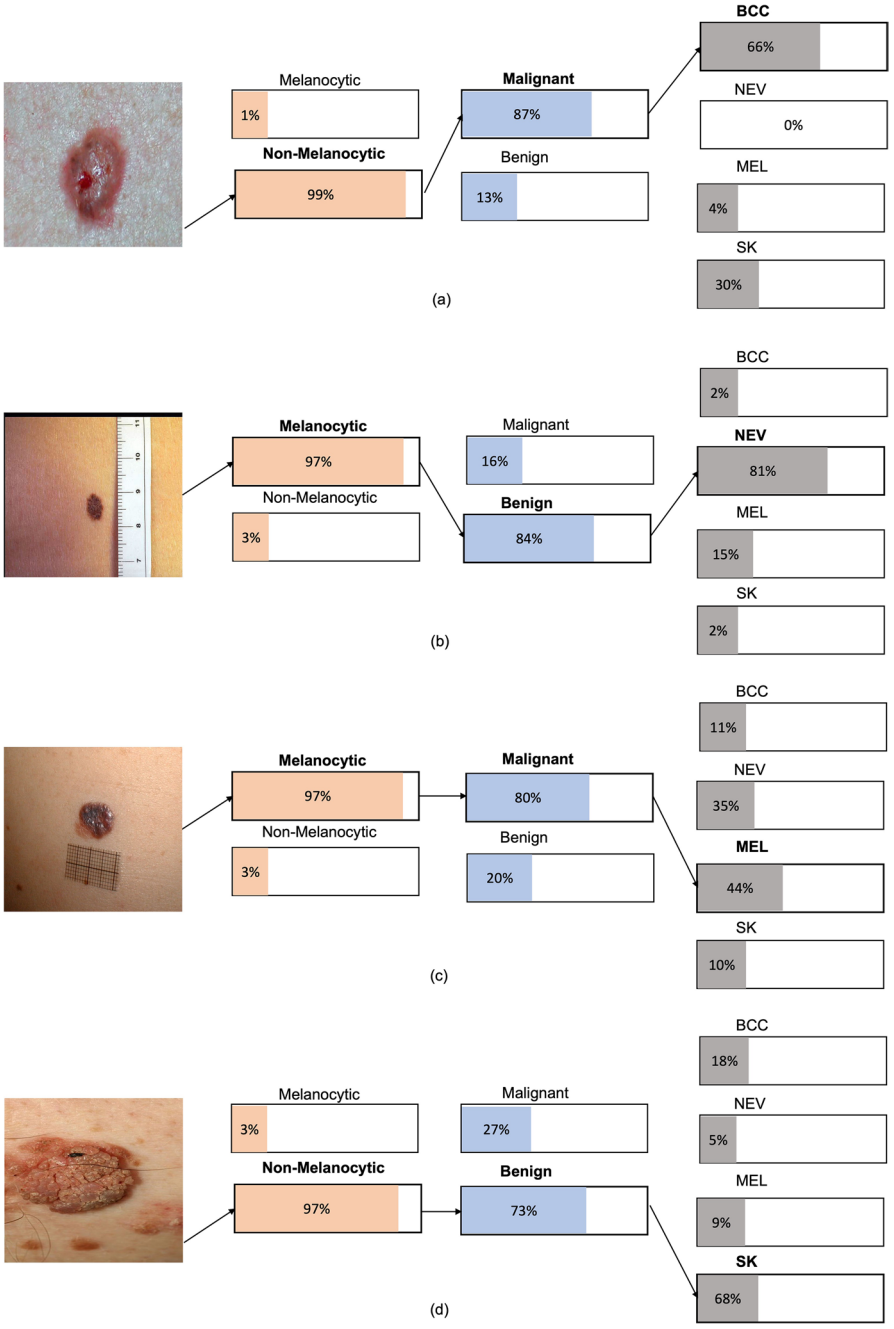

Fig. 6 Skin taxonomy interpretation of the results for BCC (**a**), NEV (**b**), MEL (**c**), and SK (**d**)

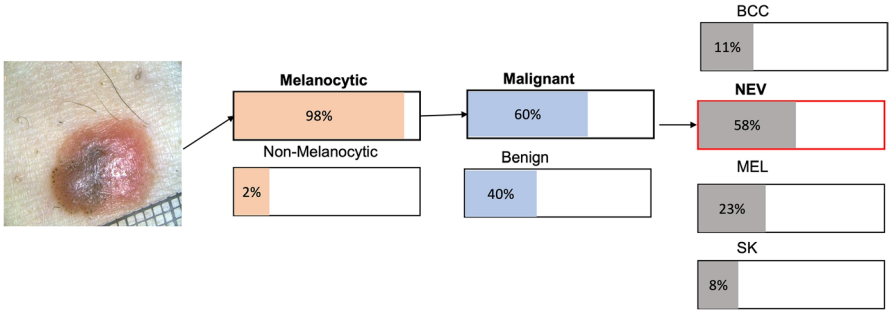

Fig. 7 Misclassified MEL as NEV

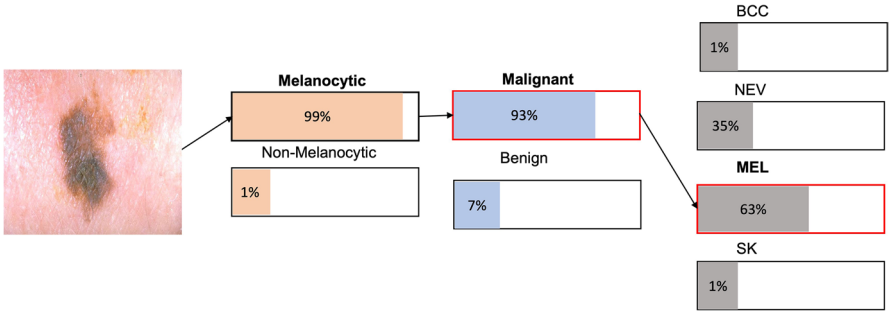

Fig. 8 Worst-case error

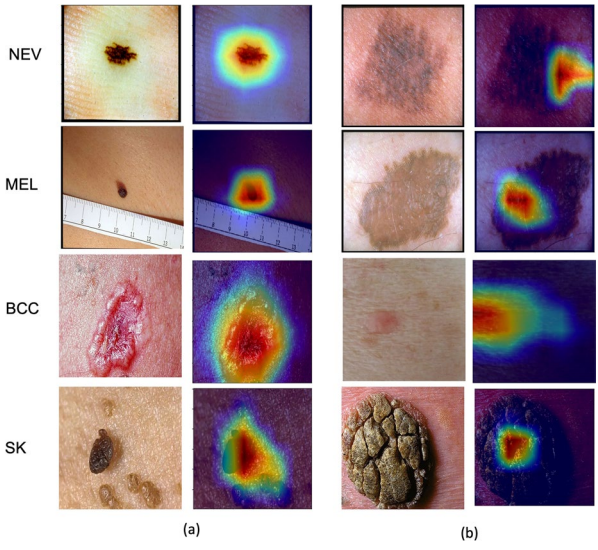

Fig. 9 GradCAM++ interpretation for the correctly classified cases (**a**) and incorrectly classified cases (**b**)

classification. It can be observed that the generated maps are consistent with the model predictions; thus, there is an overlap between the disease region and the region highlighted by the GradCAM++ in Fig. 9a for the correctly classified cases. In contrast, in Fig. 9b illustrating the misdiagnosed cases, we can notice that the regions where the model used to make a diagnosis do not align well with the disease region. This implies the GradCAM++ faithfulness in explaining the results of the model.

7 Discussion

In this work, an interpretable skin cancer diagnosis approach utilizing clinical images to learn dermatological concepts was developed and evaluated. The results showed the significant impact of the incremental architecture, the taxonomy-based loss function, and the minority class data integration on improving the model's accuracy and boosting its ability to differentiate between the classes in all taxonomy levels. In addition, the model demonstrated its ability to justify the diagnosis and discover disease misdiagnosis.

In comparison with prior works, Kawahara et al. [49] developed a DNN to diagnose skin lesions using the same dataset utilized herein our work (discussed in Sect. 3). The model of Kawahara et al. trained on clinical images had an average accuracy of 60% compared to 67% average disease accuracy achieved by our model implementing the incremental architecture, taxonomy-based loss function, and data transformation without integrating images from DermNet NZ (Table 3). In addition, our model outperforms the classification approach proposed by Ngiam et al. [76] and implemented by Kawahara et al. [49] using the same clinical images; Ngiam's model had an accuracy of 58.2% compared to 67% for our model. In terms of interpretability, Kawahara et al. utilized CAM to visualize the dominant regions of the lesion images employed for diagnosis [49].

Ge et al. [44] developed three DNNs trained and tested on 26,584 clinical images to predict 15 skin lesions (3 malignant and 12 benign conditions) and achieved average accuracies of 52.2%, 54.1%, and 59.4% for the three DNNs. On the other hand, we focused on 4 diseases as our focus is more on skin cancer detection than common skin conditions. We employed 1108 images for developing our best performing model (M_4) which achieved an average disease level accuracy of 71% and all levels' average accuracy of 78%. With respect to interpretability, Ge et al. integrated CAM with the bilinear feature pooling [77] to provide a detailed visual explanation of the important regions of skin lesion images.

Esteva et al. [58] trained an Inception V3 network on 129,450 clinical images to classify the lesion as carcinoma versus SK and melanoma versus nevi. Although the selection of the disease classification tasks was based on a skin lesion taxonomy, the authors did not consider the taxonomy in the models' implementation. In terms of accuracy, our best performing model achieved 77% accuracy in malignancy detection and 71% accuracy in diagnosing the four diseases; however, in Esteva's work, the accuracy of classifying carcinomas and SK was 72.1%, and the accuracy of diagnosing melanoma and nevi was 55.4%. Regarding interpretability, Esteva et al. employed the t-distributed stochastic

neighbor embedding [78] that visualizes high dimensional data to envision the learnt features of the DNN's last layer and thus understand the inference of the model.

Although our approach outperforms other prior works, it has some limitations. The disease level accuracy can be further improved by incorporating more clinical images in model training. In addition, various loss functions that reflect the taxonomy can be developed to investigate their impact on classification accuracy. Finally, evaluating the developed interpretability methods by general practitioners is needed to assess the impact of the implemented skin cancer diagnosis models on the performance and confidence of humans.

8 Conclusion

We presented an interpretable skin cancer diagnosis approach that employs a skin lesion taxonomy to incrementally learn dermatologic knowledge using an adapted DNN architecture. Our models were trained on clinical images as they are easily obtained by a non-specialist healthcare provider. The empirical analyses showed that the implemented taxonomy is beneficial in improving classification accuracy, understanding the rationale behind the disease diagnosis, and discovering diagnosis errors. Moreover, we employed an advanced gradient-based class activation map method that demonstrated consistent visual explanations of the diagnosis-making process. Our work is a step toward developing an interpretable rapid skin cancer diagnostic tool that can assist general practitioners to make an early diagnosis. Further long-term, large-scale validation studies are nonetheless needed to understand the usability, interpretability, and accuracy of our proposed model when employed by general practitioners in clinical settings.

Author Contribution All authors designed the analysis. E.R. performed the analysis, prepared the figures, and wrote the manuscript draft. All authors revised the manuscript.

Funding This work is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Declarations

Competing Interests The authors declare no competing interests.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A (2022) Cancer statistics. CA Cancer J Clin 72:7–33. https://doi.org/10.3322/caac.21708
2. American Cancer Society (2022) Cancer facts & figures 2022. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2022.html. Accessed 12 June 2022
3. American Academy of Dermatology Association (AAD) (2022) Skin cancer. https://www.aad.org/media/stats-skin-cancer. Accessed 12 June 2022
4. Resneck J, Kimball AB (2004) The dermatology workforce shortage. J Am Acad Dermatol 50:50–54. https://doi.org/10.1016/j.jaad.2003.07.001

5. Feng H, Berk-Krauss J, Feng PW, Stein JA (2018) Comparison of dermatologist density between urban and rural counties in the United States. JAMA Dermatol 154:1265–1271. https://doi.org/10.1001/jamadermatol.2018.3022
6. Ramsay DL, Weary PE (1996) Primary care in dermatology: whose role should it be? J Am Acad Dermatol 35:1005–1008. https://doi.org/10.1016/S0190-9622(96)90137-1
7. Moreno G, Tran H, Chia ALK, Lim A, Shumack S (2007) Prospective study to assess general practitioners' dermatological diagnostic skills in a referral setting. Australas J Dermatol 48:77–82. https://doi.org/10.1111/j.1440-0960.2007.00340.x
8. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A et al (2019) A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 111:148–154. https://doi.org/10.1016/j.ejca.2019.02.005
9. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S et al (2019) Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 119:11–17. https://doi.org/10.1016/j.ejca.2019.05.023
10. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A et al (2019) Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 113:47–54. https://doi.org/10.1016/j.ejca.2019.04.001
11. Maron RC, Weichenthal M, Utikal JS, Hekler A, Berking C et al (2019) Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Cancer 119:57–65. https://doi.org/10.1016/j.ejca.2019.06.013
12. Goyal M, Knackstedt T, Yan S, Hassanpour S (2020) Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities. Comput Biol Med 127:104065. https://doi.org/10.1016/j.compbiomed.2020.104065
13. Hauser K, Kurz A, Haggenmüller S, Maron RC, von Kalle C et al (2022) Explainable artificial intelligence in skin cancer recognition: a systematic review. Eur J Cancer 167:54–69. https://doi.org/10.1016/j.ejca.2022.02.025
14. Holzinger A (2021) The next frontier: AI we can really trust. Mach Learn Princ Pract Knowl Discov Databases ECML PKDD 2021, CCIS, vol 1524. Springer, Cham, pp 427–440. https://doi.org/10.1007/978-3-030-93736-2_33
15. Madiega T, Chahri S (2022) BRIEFING: EU legislation in progress, proposal for artificial intelligence act. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792. Accessed 12 June 2022
16 BarredoArrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
17. Holzinger A, Saranti A, Molnar C, Biecek P, Samek W (2022) Explainable AI methods - a brief overview. xxAI - beyond explain AI xxAI 2020 Lect Notes Comput Sci, vol 13200. Springer, Cham, pp 13–38. https://doi.org/10.1007/978-3-031-04083-2_2
18. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P (2022) Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput Biol Med 140:105111. https://doi.org/10.1016/j.compbiomed.2021.105111
19. Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI : a review of machine learning interpretability methods. MDPI Entropy 23:e23010018. https://doi.org/10.3390/e23010018
20. Joshi G, Walambe R, Kotecha K (2021) A review on explainability in multimodal deep neural nets. IEEE Access 9:59800–59821. https://doi.org/10.1109/ACCESS.2021.3070212.A
21. Fuhrman JD, Gorre N, Giger ML, Hu Q, Li H (2022) A review of explainable and interpretable AI with applications in COVID-19 imaging. Med Phys 49:1–14. https://doi.org/10.1002/mp.15359
22. Barata C, Marques JS (2019) Deep learning for skin cancer diagnosis with hierarchical architectures. IEEE 16th Int Symp Biomed Imaging 2019:841–845. https://doi.org/10.1109/ISBI.2019.8759561
23. Xie Y, Zhang J, Xia Y, Shen C (2020) A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Trans Med Imaging 39:2482–2493. https://doi.org/10.1109/TMI.2020.2972964
24. Nachbar F, Stolz W, Merkle T, Cognetta AB, Vogt T, Landthaler M, Bilek P, Braun-Falco O, Plewig G (1994) The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. J Am Acad Dermatol 30:551–559. https://doi.org/10.1016/S0190-9622(94)70061-3

25. Blahnik V, Schindelbeck O (2021) Smartphone imaging technology and its applications. Adv Opt Technol 10:145–232. https://doi.org/10.1515/aot-2021-0023
26 Fee J, McGrady F, Rosendahl C, Hart N (2019) Dermoscopy use in primary care: a scoping review. Dermatol Pract Concept 9(2):98–104. https://doi.org/10.5826/dpc.0902a04
27. Barata C, Santiago C (2021) Improving the explainability of skin cancer diagnosis using CBIR. Med Image Comput Comput Assist Interv – MICCAI 2021 Lect Notes Comput Sci, vol 12903. Springer, Cham, pp 550–559. https://doi.org/10.1007/978-3-030-87199-4_52
28. Codella NCF, Lin CC, Halpern A, Hind M, Feris R et al (2018) Collaborative human-AI (CHAI): evidence-based interpretable melanoma classification in dermoscopic images. MLCN DLF IMIMIC 2018 Lect Notes Comput Sci, vol 11038. Springer, Cham, pp 97–105. https://doi.org/10.1007/978-3-030-02628-8_11
29. Abbasi NR, Shaw HM, Rigel DS, Friedman RJ, Mccarthy WH et al (2004) Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. JAMA - J Am Med Assoc 292:2771–2776
30. Chowdhury T, Bajwa ARS, Chakraborti T, Rittscher J, Pal U (2021) Exploring the correlation between deep learned and clinical features. Med Image Underst Anal MIUA 2021 Lect Notes Comput Sci, vol 12722. Springer, Cham, pp 3–17. https://doi.org/10.1007/978-3-030-80432-9_1
31. Stieler F, Rabe F, Bauer B (2021) Towards domain-specific explainable AI: model interpretation of a skin image classifier using a human approach. IEEE Comput Soc Conf Comput Vis Pattern Recognit Work 2021. pp 1802–1809. https://doi.org/10.1109/CVPRW53098.2021.00199
32. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. The 2016 conference of the North American chapter of the association for computational linguistics: demonstrations 2016. pp 97–101. https://doi.org/10.18653/v1/n16-3020
33. B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, (2018) Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV), in: 35th Int Conf Mach Learn ICML 2018. 6:4186–4195
34. Lucieri A, Bajwa MN, Braun SA, Malik MI, Dengel A, Ahmed S (2022) ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. Comput Methods Programs Biomed 215:106620. https://doi.org/10.1016/j.cmpb.2022.106620
35. Tschandl P, Rosendahl C, Kittler H (2012) Accuracy of the first step of the dermatoscopic 2-step algorithm for pigmented skin lesions. Dermatol Pract Concept 2:43–49. https://doi.org/10.5826/dpc.0203a08
36. Barata C, Celebi ME, Marques JS (2021) Explainable skin lesion diagnosis using taxonomies. Pattern Recognit 110:107413. https://doi.org/10.1016/j.patcog.2020.107413
37. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016. pp 2921–2929. https://doi.org/10.1109/CVPR.2016.319
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D et al (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. IEEE Int Conf Comput Vis 2017. pp 618–626. https://doi.org/10.1109/ICCV.2017.74
39. Yang J, Xie F, Fan H, Jiang Z, Liu J (2018) Classification for dermoscopy images using convolutional neural networks based on region average pooling. IEEE Access 6:65130–65138. https://doi.org/10.1109/ACCESS.2018.2877587
40. Wei L, Ding K, Hu H (2020) Automatic skin cancer detection in dermoscopy images based on ensemble lightweight deep learning network. IEEE Access 8:99633–99647. https://doi.org/10.1109/ACCESS.2020.2997710
41. Zunair H, Ben Hamza A (2020) Melanoma detection using adversarial training and deep transfer learning. Phys Med Biol 65:135005 https://doi.org/10.1088/1361-6560/ab86d3
42. Li W, Zhuang J, Wang R, Zhang J (2020) Fusing metadata and dermoscopy images for skin disease diagnosis. IEEE 17th Int Symp Biomed Imaging 2020. pp 1996–2000
43. Nunnari F, Kadir MA, Sonntag D (2021) On the overlap between Grad-CAM saliency maps and explainable visual features in skin cancer images. Mach Learn Knowl Extr, vol 12844. Springer, Cham, pp 241–253. https://doi.org/10.1007/978-3-030-84060-0_16
44. Ge Z, Demyanov S, Chakravorty R, Bowling A, Garnavi R (2017) Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. Med Image Comput Comput Assist Interv – MICCAI 2017 Lect Notes Comput Sci, vol 10435. pp 250–258. https://doi.org/10.1007/978-3-319-66179-7_29
45. Lin TY, Roychowdhury A, Maji S (2015) Bilinear CNN models for fine-grained visual recognition. IEEE Int Conf Comput Vis 2015. pp 1449–1457. https://doi.org/10.1109/ICCV.2015.170

46. Han SS, Kim MS, Lim W, Park GH, Park I et al (2018) Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 138:1529–1538. https://doi.org/10.1016/j.jid.2018.01.028
47. Pfau J, Young AT, Wei ML, Keiser MJ (2019) Global saliency: aggregating saliency maps to assess dataset artefact bias. Machine Learning for Health (ML4H) Workshop at NeurIPS 2019. pp 1–9.https://doi.org/10.48550/arXiv.1910.07604
48. Gupta A, Arora S (2019) A simple saliency method that passes the sanity checks. ArXiv 2019. pp 1–11.https://doi.org/10.48550/arXiv.1905.12152
49. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G (2019) Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Heal Informatics 23:538–546. https://doi.org/10.1109/JBHI.2018.2824327
50. Argenziano G, Fabbrocini G, Carli P, De Giorgi V, Sammarco E et al (1998) Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arch Dermatol 134:1563–1570. https://doi.org/10.1001/archderm.134.12.1563
51. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. The 34th Int Conf Mach Learn, vol 70. pp 3319–3328
52. Liu Y, Jain A, Eng C, Way DH, Lee K et al (2020) A deep learning system for differential diagnosis of skin diseases. Nat Med 26:900–908. https://doi.org/10.1038/s41591-020-0842-3
53. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M (2017) SmoothGrad: removing noise by adding noise. ArXiv 2017. pp 1–10. https://doi.org/10.48550/arXiv.1706.03825
54. Singh N, Lee K, Coz D, Angermueller C, Huang S et al (2020) Agreement between saliency maps and human-labeled regions of interest: applications to skin disease classification. IEEE Comput Soc Conf Comput Vis Pattern Recognit Work 2020. pp 3172–3181. https://doi.org/10.1109/CVPRW50498.2020.00376
55. Seven point Criteria Evaluation Database (2019). https://derm.cs.sfu.ca/Welcome.html. Accessed 20 May 2022
56. Zhu R, Guo Y, Xue JH (2020) Adjusting the imbalance ratio by the dimensionality of imbalanced data. Pattern Recognit Lett 133:217–223. https://doi.org/10.1016/j.patrec.2020.03.004
57. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016. pp 2818–2826. https://doi.org/10.1109/CVPR.2016.308
58. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056
59. Gordon-Rodriguez E, Loaiza-Ganem G, Pleiss G, Cunningham JP (2020) Uses and abuses of the cross-entropy loss: case studies in modern deep learning. Mach Learn Res ICBINB, NeurIPS, PMLR 37:1–10. https://proceedings.mlr.press/v137/gordon-rodriguez20a.html. Accessed 12 June 2022
60. Molnar C (2022) Neural networks interpretation. Interpretable Machine Learning: a Guide for Making Black Box Model Explainable Second edition chapter 10:444–473
61. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. IEEE Winter Conf Appl Comput Vision, WACV 2018. pp 839–847. https://doi.org/10.1109/WACV.2018.00097
62. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6:6–27. https://doi.org/10.1186/s40537-019-0192-5
63. Sugino T, Kawase T, Onogi S, Kin T, Saito N et al (2021) Loss weightings for improving imbalanced brain structure segmentation using fully convolutional networks. MDPI Healthc 9(8):938. https://doi.org/10.3390/healthcare9080938
64. Cui Y, Jia M, Lin TY, Song Y, Belongie S (2019) Class-balanced loss based on effective number of samples. IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019. pp 9260–9269. https://doi.org/10.1109/CVPR.2019.00949
65. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6:60. https://doi.org/10.1186/s40537-019-0197-0
66. DermNet NZ, (2013). https://dermnetnz.org/ Accessed 7 Feb 2022
67. Kumar R (2019) Cross-validation and model selection. Machine learning quick reference: quick and essential machine learning hacks for training smart data models. Packet Publishing, pp 27–29.
68. Cuemath Z Test, (2016). https://www.cuemath.com/data/z-test/ Accessed 24 Nov 2022

69. Tan C, Sun F, Kong T, Zhang W, Yang C et al (2018) A survey on deep transfer learning. Artificial Neural Networks and Machine Learning – ICANN 2018. ICANN 2018, Lecture Notes in Computer Science, vol 11141. pp 70–279. https://doi.org/10.1007/978-3-030-01424-7_27
70. Russakovsky O, Deng J, Su H, Krause J, Satheesh S et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115:211–252. https://doi.org/10.1007/s11263-015-0816-y
71. Lin M, Chen Q, Yan S (2014) Network in network. The 2nd Int Conf on Learn Rep ICLR 2014. pp 1–10. https://doi.org/10.48550/arXiv.1312.4400
72. Ruder S (2017) An overview of gradient descent optimization algorithms. ArXiv 2017. pp 1–14. https://doi.org/10.48550/arXiv.1609.04747
73. Chollet F (2015) Keras. https://github.com/fchollet/keras. Accessed 24 Apr 2022
74. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z et al (2015) TensorFlow: large-scale machine learning on heterogeneous distributed systems. https://tensorflow.org. Accessed 24 Apr 2022
75. Hossin M, Sulaiman M (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5:01–11. https://doi.org/10.5121/ijdkp.2015.5201
76. Ngiam J, Chen Z, Koh PW, Ng AY (2011) Learning deep energy models. The 28th Int Conf Mach Learn ICML 2011. pp 1105–1112
77. Gao Z, Wu Y, Zhang X, Dai J, Jia Y, et al (2020) Revisiting bilinear pooling: a coding perspective. The 34th AAAI Conf Artif Intell 2020. pp 3954–3961. https://doi.org/10.1609/aaai.v34i04.5811
78. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. Mach Learn Res 9:2579–2605. https://doi.org/10.1007/s10479-011-0841-3