Test-Time Adaptation for Skin Lesion Classification

THOMAS BOSTELAAR, University of Twente, The Netherlands

Deep learning models have shown good potential in the skin lesion classification task. In the case of a domain shift, where the test data comes from a different distribution than the training data, deep learning models struggle to perform. Dermoscopy images are taken under a wide range of different circumstances, and as a result distribution shifts can exist between different data sources. In this paper standard deep learning models are combined with test-time adaptation techniques to adapt to such domain shifts in the setting of binary skin lesion classification. By using a dataset split based on visual properties, models are confronted with a domain shift. Several test-time adaptation techniques are used in an online setup and compared to the unadapted version of the model. The performance of the models are analysed and attention maps are used to better understand some of the performances. Based on the results, it cannot be concluded that test-time adaptation offer a stable improvement over standard deep learning models.

Additional Key Words and Phrases: Skin Lesion Classification, Test-Time Adaptation, Deep Learning

1 INTRODUCTION

Skin cancer opposes a serious problem to humankind, with Melanoma alone accounting for 4% of all new cancer diagnoses in EU-27 countries in 2020, causing 1.3% of the deaths [1]. Using dermoscopy, also known as skin surface microscopy, skin lesions can be examined in greater detail. Deep neural networks have shown to deliver good results in classifying skin cancer diseases [2] and can be utilised in computer-aided diagnosis (CAD) to deliver accurate diagnosis. The drawback of deep learning methods is that they do not generalize well. Real world data can differ from the training data and as a result there may be a distribution shift between training and the test data. Deep learning models tend have performance drops when confronted with distribution shifts between training and the test data. Dermoscopic images may have different visual properties based on the clinic or patient. It is essential that a model generalize well to these different settings. Additionally, domain adaptation techniques can be used to adapt a model to new domains. A form of domain adaptation called test-time adaptation (TTA) is able to utilize unlabeled test data during test time to make changes to the underlying model.

In this paper the research will be focused on TTA in a binary classification setting for the skin lesion classification task. Using TTA the model is able to leverage additional knowledge during test time. The aim is to investigate if TTA methods can be applied to a skin lesion dataset with domains based on the presence of visual artifacts.

Existing research on skin lesion classification often focuses on domain adaptation using source data or domain generalization to deal with domain shifts. In this paper the focus is on TTA a sub-field of domain adaptation, using only a pretrained model and unlabeled test data.

The main objective of this research is summarized in the following research questions:

- How does test-time adaptation affect the performance of skin lesion classification on domain-shifted target data?
- What is the effect of test-time adaption on bias in skin lesion classification?

The first and main research question is addressed by comparing two base models against their adapted versions using standard classification metrics. The second research question is answered by visualizing and comparing the class activation mappings (CAMs) on a set of samples between the base models and the adapted models. In the context of the second research question bias is here defined as a model making incorrect assumptions about the target data. A model might base its predictions on artifacts present in the image and not on the skin lesion itself. Using CAM analysis a deeper understanding is gained of the underlying assumptions of the models before and after adaptation.

The paper is structured as follows: Relevant research related to this paper is outlined in the related works section. The methodology sections provides a detailed description of the experiments carried out the answer the research questions. The results are reported and discussed in the result section. The paper is finished with a conclusion section.

2 RELATED WORK

2.1 Computer vision for skin lesion analysis

The two main tasks in computer vision for skin lesions analysis are segmentation and classification. Segmentation aims to extract the region of the skin lesion, classification on the other hand is used to classify a skin lesion into specific categories.[3] This paper focuses on a binary classification task distinguishing between benign and malignant skin lesions. Two subfields that are discussed into further detail:

- *Unsupervised Domain Adaptation (UDA)* is a term used for the adaptation of a model using unlabeled test data, usually in combination with the labeled source/training data.
- *Domain Generalization (DG)* aims to train a model to generalize well to unseen domains often using multiple domains in the training set.

2.2 Unsupervised Domain Adaptation

Several works examine the use of UDA methods to improve skin lesion classification. [4] group images into domains based on metadata. An UDA method called Domain Adverserial Neural Network (DANN) [5] was shown to be effective in this setup. Using the same methodology a benchmark study was done comparing 8 state-of-the art UDA methods [6]. All UDA methods showed an improve in performance for most datasets compared to the unadapted model. The

TScIT 40, 2024, Enschede, The Netherlands

effectiveness of UDA across different skin lesion datasets was examined in [7] and shown to be superior to base models for single-source binary classification.

2.3 Domain Generalization

Classification models may rely on spurious correlations present in skin lesion datasets. [8] A set of 7 visual artifacts (dark corners, hair, gel border, ruler, ink markings, gel bubbles and patches) that could introduce bias are annotated in [9]. Adversarial training using a method called Learning Not To Learn (LTNL) was used in attempt to debias models, but the issue was proven too complex. In another work [10], the same artefacts were used to create the so-called trap sets. The approach that was taken in this work was a combination of domain generalization and test-time feature selection. To prevent models from overfitting to artifacts present in images a domain generalization approach is taken in [11]. The authors propose an environment aware prompt vision transformer (EVPT), which uses domain-specific and cross-domain knowledge to improve domain generalization.

2.4 Test-time adaptation

Test-time adaptation (TTA) is a type of domain adaptation where a pretrained model is adapted to unlabeled target data without access to the source data. The adaptation takes place during test time. The advantage of TTA over standard unsupervised domain adaptation (UDA) is that TTA does not need the source data, but only needs the pretrained model and the unlabeled test data. privacy regulations. A recent survey on TTA categories TTA methods into three separate categories[12] . *Source-Free Domain Adaptation* adapts the model using the entire dataset, all batches are used before making predictions, *Test-Time Batch Adaptation* uses one or a few instances, but the predictions of each batch are independent and *Online Test-Time Adaptation* adapts the model batch-by-batch, but can use information learned from previous batches.

Test-time template adjuster (T3A) aims to adjust the output layer of the deep neural network during test-time [13]. The model is split into the featurizer and the classifier, where the classifier represents the output layer and the featurizer the rest of the network. T3A first uses the model to obtain the pseudo labels $\hat{y}$. A support set for each class is created and filtered on prediction entropy. The final predictions are made using nearest centroid classification on the centroids of the support set.

Sharpness-**a**ware and **r**eliable entropy minimization (SAR) lists multiple scenarios which might negatively impact TTA methods: mixed distribution shifts, small batch sizes and label distribution shifts / label imbalance [14] . The paper states that group or layer norm based layers provide more stable TTA result than batch norm layers. The reason for this is that it is difficult to estimate BN statistics in a wild/real life setting. Samples with low entropy are removed to prevent large gradients which can lead to a model collapse. SAM (Sharpness-Aware Minimization) is used as an optimizer to make the model go towards a flat entropy surface. [15]

Source-**H**yp**O**thesis **T**ransfer (SHOT) freezes the classifier module and aims to learn the feature extraction module [16]. SHOT takes the opposite approach of T3A, where the aim is to learn the classifier module. A self-supervised pseudo-labeling strategy is proposed. SHOT starts by calculating the centroids of the target classes. A nearest centroid classifier is used based on the cosine distance to get the pseudolabels for the unlabeled test data. Based on the pseudolabels the target centroids are calculated. The proposed setting is an offline setting where the model is adapted using the whole dataset over multiple epochs. T3A provides an adaptation of SHOT to an online setting,

3 METHODOLOGY

3.1 Dataset

The dataset is supplement by [11] is based on the ISIC 2019 [17] training dataset. The authors used a classifier to separate the ISIC dataset into 5 separate domains based on visual properties. The dataset contains the following domains:

- *clean*: images containing only a clear skin and the skin lesion
- *hair*: contains hairs in part of the images
- *dark corner*: contains the corners of a camera lens
- *gel bubble*: contains gel bubbles that are applied during dermoscopy
- *ruler*: contains either a physical ruler or a ruler from software in the images.

The dataset distribution is shown table 1. It is clear from the distribution that the dataset is imbalanced, which is addressed during preprocessing. Fig. 2 shows samples for each domain in the dataset. Before training the models, oversampling is applied to the malignant class which is the minority class. A simple approach is taken by increasing the size of the malignant class threefold. The oversampling should prevent the model from simply leaning towards the majority class and predict most samples as benign. To prevent the model from overfitting on duplicated examples, data augmentation is implemented. Several transformations are applied including flipping, changing brightness and contrast and blurring. The augmentations are applied to the entire training dataset. The aim of this research is not to maximize the performance of the base models, but to ensure the base models provide stable results oversampling and augmentation is implemented. Data augmentation may negatively impact the performance of TTA [18].

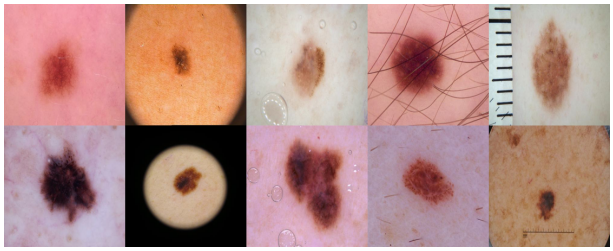

Fig. 1. Samples from the 5 categories: Clean, Dark Corner, Gel Bubble, Hair, Ruler. Top: Benign, Bottom: Malignant

Category	Benign	Malignant	Total
Clean	2385	411	2796
Dark Corner	1747	604	2351
Gel Bubble	1211	429	1640
Hair	4267	617	4884
Ruler	509	163	672
All	10119	2224	12343

Table 1. Dataset distribution

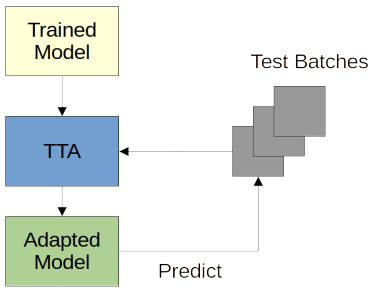

Fig. 2. TTA Pipeline

3.2 Training models

The objective of the experiments is to simulate a domain shift by testing the model on a domain that contains data not seen in the training data and evaluate how TTA affects the model. Since the TTA methods operate during the testing phase, a separate model is first trained which acts as a backbone for the test-time adaptation. For the backbone two different models are selected. The first choice is a ResNet50 which is a common choice for image classification. As a second choice the small version of a Data-Efficient Image Transformer (DeiT) model is used. DeiT models can be trained in two ways, in the classic way with a linear layer on top of the final hidden state of CLS token or using a linear layer on both the CLS token and the distillation token. The average of the predictions can then be taken. The simple approach is taken by training the model in the classic way by only placing one linear layer. DeiT III is used which modifies the training process of the original DeiT [19]. The pretrained ResNet50 and DeiT models are finetuned for the skin lesion classification problem. In a setup with limited computional power available, finetuning is done for only for 10 epochs with the learning rate set experimentally to 1e-4 and using adam as an optimizer. Google Colab provides a free GPU and is used to carry out the experiments.

3.3 Validation Metrics

The evaluation is done on data that is imbalanced. Accuracy in itself is not a good metric for imbalanced datasets. Recall, precision and the f1-score provide a better understanding of the classification performance. The recall, precision and f1-score are reported for the malignant class. The f1-score and the accuracy are considered as the main metrics, with recall and precision giving more insights into the f1-score.

3.4 Implementation Details

A leave-one-out cross validation is done on the five domains to evaluate the models. The base models are trained on four domains and evaluated on the domain that is excluded from the training set. Three different TTA methods are added separately as an extra layer on top of the baseline which allows for changes during test time. The TTA methods that are evaluated are SHOT [16], T3A [13] and SAR [14]. The average values across the five domain obtained using leave-one-out cross validation are reported with their standard deviations in table 2. Accuracy, recall, precision and f1-score are reported in the appendix in tables 3, 4,5 and 6. Class activation mapping (CAM) techniques can help to obtain a better understanding of the decision making process deep learning models. Score-CAM [20], a gradient-free CAM method is used as the CAM method of choice. The TTA methods are used in an online setup, therefore the model is adapted for every batch in the dataset. T3A only adapts the linear layer/the output layer, which makes it not suitable for analysis. SHOT delivers unstable results. The TTA method that is used for the visualization is therefore SAR. Because the model is evolving throughout the dataset, this brings a challenge for analyzing the attention map. The simple approach is taken of using the adapted model after it has passed through the entire test set, which should give a good approximation of an adapted model. The CAM analysis is done by comparing the base models ResNet50 and DeiT with the adapted ResNet50 by SAR.

The implementation of SHOT by T3A is used. SAR was added to the code base as an additional method. During evaluation SHOT results were unstable, therefore the parameter β was run for the values [0.3,0.6,0.9] with α experimentally set to 0.1. A β of 0.6 provided the most stable results and therefore the evaluation of SHOT is reported using this value. SAR has a parameter used for selecting samples based on an entropy threshold which is set to the default value of $0.4 \cdot \ln 1000$. For T3A the a value of fixed value 5 is used for the filter-K parameter, which is used to filter support sets based on entropy.

4 RESULTS

4.1 Results

The summary of the Leave-one-out cross validation (LOOCV) is reported in Table 2, with the average values reported for each metric. The full results of the LOOCV for each test domain are displayed in the appendix in tables 3,4, 5 and 6.
Figure 3 shows the CAM for a set of examples from the gel bubble and dark corner across the two base models and one adapted model.

4.2 Discussion

The results in Table 2 show that TTA methods do not offer a significant improvement to the base models under the chosen classification metrics. The combination of DeiT and SAR shows small improvement in both accuracy and f1-score. However, the full results in tables 4,5 and 6 show that the DeiT base model was often unchanged by SAR, therefore no strong conclusion can be drawn from these

Table 2. Results obtained by leave-one-out cross validation on 5 domains. Base models ResNet50 and DeiT-S are compared with their adapted versions.

Models	Accuracy	Recall	Precision	F1
Resnet50	**0.821**±0.032	0.601±0.096	0.566±0.102	**0.577**±0.079
+T3A	0.761±0.046	0.728±0.066	0.456±0.099	0.555±0.083
+SAR	0.774±0.037	0.707±0.054	0.484±0.132	0.560±0.094
+SHOT	0.680±0.121	0.283±0.310	0.307±0.070	0.187±0.115
DeiT-S	0.833±0.021	0.659±0.035	0.586±0.097	0.616±0.063
+T3A	0.824±0.014	0.682±0.016	0.564±0.107	0.612±0.067
+SAR	**0.834**±0.022	0.657±0.038	0.591±0.096	**0.618**±0.062
+SHOT	0.679±0.045	0.734±0.176	0.357±0.102	0.475±0.120

results. Overall SAR performs best out of the TTA method both on accuracy and f1-score. T3A results are stable but show a decline over the baselines. Similar to SAR, T3A shows no big changes when applied to the DeiT base model, which shows a further investigation into the hyperparameters might be necessary. SHOT results are very unstable with often low values for precision and recall. The DeiT model shows an improvement the ResNet50 model, showing the potential for transformer-based architectures. As outlined by SAR architectures using layer normalization are better suited for TTA as opposed to the usage of batch normalization.[14]

The CAMs for several benign and malignant examples are visualized in 3. From this limited set of samples no improvement is seen by SAR. In one case SAR even shifts the attention away from the skin lesion towards a gel bubble. This example shows SAR can introduce bias instead of debiasing the model.

The conclusions from the results are that TTA does not positively impact skin lesion classification in the setup of this paper. Furthermore, based on a limited set of examples, it can be concluded that TTA fails to remove bias from model and can do the reverse by introducing bias.

Some of the reasons that can cause TTA to fail are outlined SAR paper, relevant are mixed distributions shifts and imbalanced test data [14]. The domains in the dataset contain images from other domains and are not perfectly separated. Although SAR aims to solve some of these problems, mixed distribution shifts and imbalanced data make it a difficult setup for TTA methods. The biggest problem might be outlined by the Score-CAM analysis, TTA can shift the attention towards artifacts present in the image. The results present evidence that TTA might not be well suited for the skin lesion classification problem.

4.3 Future Work

The research in this paper is done in a limited setup combining TTA methods with skin lesion classification. Only a small subset consisting of three TTA methods are evaluated. To obtain a better understanding of how TTA methods perform in this scenario, a wider range of TTA methods with different approaches should be evaluated. In the current setup the domains are all taken from the same dataset. The effects of TTA could be further explored by testing on out-of-sample data from a different dataset. Since TTA did not deliver significant improvements, futher research might be better spent on building robust models using techniques related to domain generalization.

5 CONCLUSIONS

In this paper test-time adaptation methods are applied to a binary skin lesion classification problem. After using leave-one-out cross validation to compare the performance of TTA methods against base models, no significant increase of performance was found. Furthermore, the analysis of TTA using CAM visualization show no improvement over the base model. Compared to benchmark datasets that TTA methods are tested on, the setup of this research is a challenging one. Mixed domain shifts with imbalanced label distributions make it difficult for the models to adapt. Under the setup of this paper test-time adaptation methods did not provide any significant improvements to the performance of skin lesion classification under the chosen metrics. Additionally analysis using Score-CAM did not show signs of improvement. The evaluation of both base and adapted models show that TTA might not be well suited for the skin lesion classification task.

REFERENCES

[1] Ecis, 2020.

[2] Kuprel B. Novoa R. et al. Esteva, A. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[3] Zahra Mirikharaji, Kumar Abhishek, Alceu Bissoto, Catarina Barata, Sandra Avila, Eduardo Valle, M. Emre Celebi, and Ghassan Hamarneh. A survey on deep learning for skin lesion segmentation. *Medical Image Analysis*, 88:102863, 2023.

[4] Katharina Fogelberg, Sireesha Chamarthi, Roman C. Maron, Julia Niebling, and Titus J. Brinker. Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. *New Biotechnology*, 76:106–117, 2023.

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016.

[6] Sireesha Chamarthi, Katharina Fogelberg, Roman C. Maron, Titus J. Brinker, and Julia Niebling. Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods on dermoscopic images, 2023.

[7] Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Can domain adaptation improve accuracy and fairness of skin lesion classification?, 2023.

[8] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de)constructing bias on skin lesion datasets, 2019.

[9] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast, 2020.

[10] Alceu Bissoto, Catarina Barata, Eduardo Valle, and Sandra Avila. Artifact-based domain generalization of skin lesion models, 2022.

[11] Siyuan Yan, Chi Liu, Zhen Yu, Lie Ju, Dwarikanath Mahapatra, Victoria Mar, Monika Janda, Peter Soyer, and Zongyuan Ge. Epvt: Environment-aware prompt vision transformer for domain generalization in skin lesion recognition, 2023.

[12] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts, 2023.

[13] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[14] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023.

[15] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2021.

[16] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6028–6039. PMLR, 13–18 Jul 2020.

[17] Marc Combalia, Noel Codella, Veronica Rotemberg, Cristina Carrera, Stephen Dusza, David Gutman, Brian Helba, Harald Kittler, Nicholas Kurtansky, Konstantinos Liopyris, Michael Marchetti, Sebastian Podlipnik, Susana Puig, Christoph

Rinner, Philipp Tschandl, Jochen Weber, Allan Halpern, and Josep Malvehy. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 international skin imaging collaboration grand challenge. *The Lancet Digital Health*, 4:e330–e339, 05 2022.

[18] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation, 2023.

[19] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit, 2022.

[20] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020.

A APPENDIX

Table 3. Accuracy computed using a leave-one-out procedure, with the domain columns representing the domain that was used as a test domain.

Models	Clean	Gel Bubble	Hair	Dark Corner	Ruler	Avg.	Stdev
Resnet50	0.869	0.813	0.839	0.810	0.772	0.821	0.032
+T3A	0.819	0.779	0.757	0.771	0.680	0.761	0.046
+SAR	0.792	0.763	0.715	0.774	0.827	0.774	0.037
+SHOT	0.842	0.718	0.743	0.606	0.489	0.680	0.121
DeiT-S	0.86	0.826	0.849	0.831	0.798	0.833	0.021
+T3A	0.836	0.828	0.838	0.822	0.798	0.824	0.014
+SAR	0.864	0.826	0.849	0.832	0.798	0.834	0.022
+SHOT	0.614	0.693	0.737	0.641	0.711	0.679	0.045

Table 4. Recall computed using a leave-one-out procedure, with the domain columns representing the domain that was used as a test domain.

Models	Clean	Gel Bubble	Hair	Dark Corner	Ruler	Avg.	Stdev
ResNet50	0.550	0.709	0.488	0.536	0.724	0.601	0.096
+T3A	0.674	0.802	0.681	0.667	0.816	0.728	0.066
+SAR	0.691	0.781	0.76	0.649	0.656	0.707	0.054
+SHOT	0.029	0.044	0.104	0.394	0.844	0.283	0.310
DeiT-S	0.603	0.685	0.645	0.656	0.706	0.659	0.035
+T3A	0.669	0.678	0.671	0.679	0.712	0.682	0.016
+SAR	0.594	0.685	0.645	0.656	0.706	0.657	0.038
+SHOT	0.871	0.772	0.389	0.806	0.834	0.734	0.176

Table 5. Precision computed using a leave-one-out procedure, with the domain columns representing the domain that was used as a test domain.

Models	Clean	Gel Bubble	Hair	Dark Corner	Ruler	Avg.	Stdev
ResNet50	0.581	0.652	0.395	0.679	0.522	0.566	0.102
+T3A	0.438	0.57	0.3	0.554	0.418	0.456	0.099
+SAR	0.393	0.548	0.275	0.562	0.641	0.484	0.132
+SHOT	0.316	0.352	0.386	0.302	0.179	0.307	0.070
DeiT-S	0.539	0.693	0.44	0.692	0.567	0.586	0.097
+T3A	0.474	0.7	0.418	0.661	0.566	0.564	0.107
+SAR	0.556	0.693	0.44	0.697	0.567	0.591	0.096
+SHOT	0.261	0.46	0.21	0.405	0.449	0.357	0.102

Table 6. F1-score computed using a leave-one-out procedure, with the domain columns representing the domain that was used as a test domain.

Models	Clean	Gel Bubble	Hair	Dark Corner	Ruler	Avg.	Stdev
ResNet50	0.565	0.679	0.437	0.599	0.607	0.577	0.079
+T3A	0.531	0.667	0.417	0.605	0.553	0.555	0.083
+SAR	0.501	0.644	0.404	0.602	0.648	0.560	0.094
+SHOT	0.053	0.079	0.164	0.342	0.295	0.187	0.115
DeiT-S	0.569	0.689	0.523	0.673	0.628	0.616	0.063
+T3A	0.555	0.689	0.515	0.67	0.63	0.612	0.067
+SAR	0.574	0.689	0.523	0.676	0.628	0.618	0.062
+SHOT	0.402	0.576	0.273	0.539	0.584	0.475	0.120

B SCORE-CAM ANALYSIS

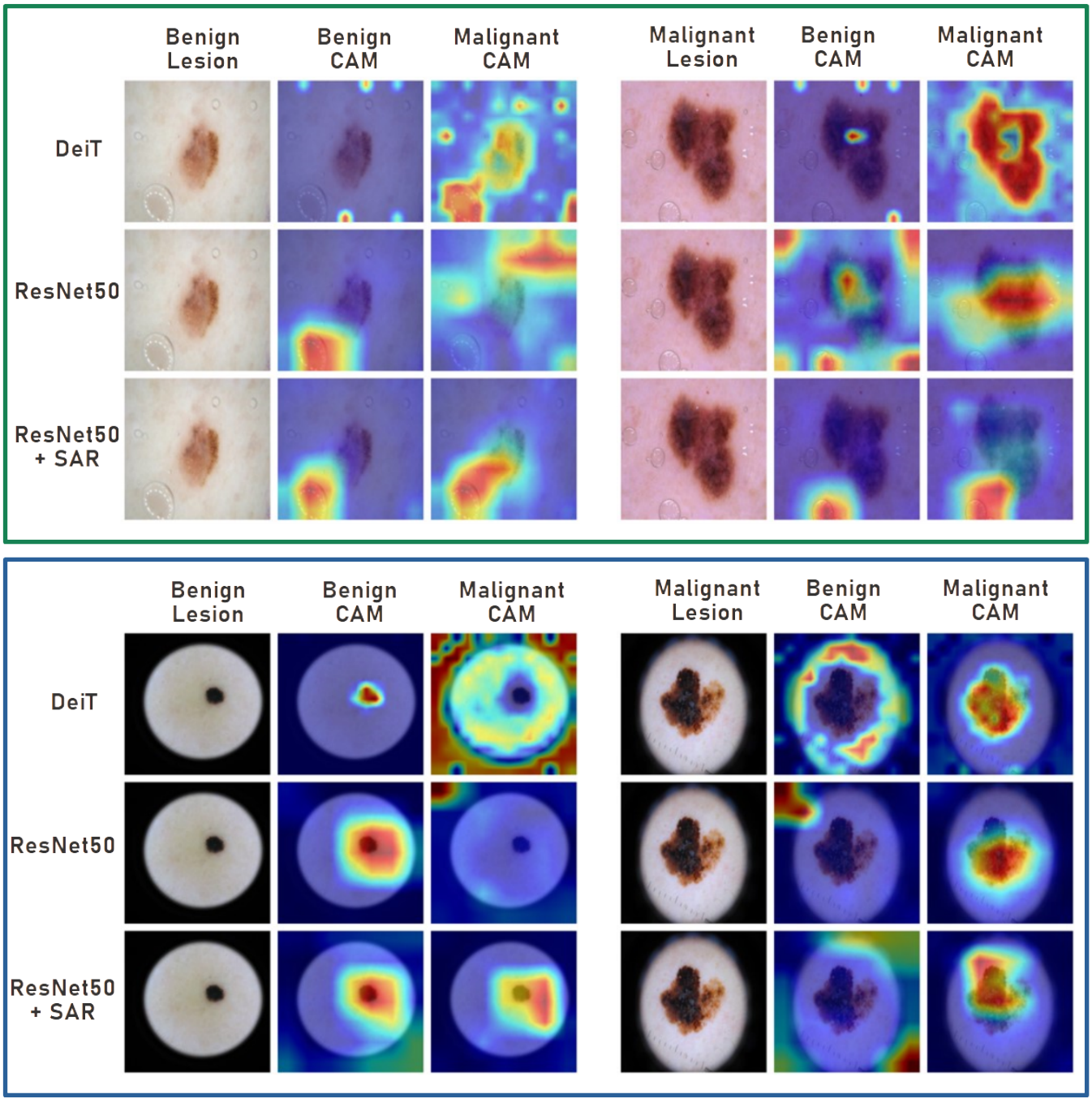

Fig. 3. Score-CAM visualisations for benign and malignant samples from the gel bubble and dark corner domain.