Explainable Skin Lesion Classification with Multitask Learning

Chih-Hao Liu and Sheng-Lung Huang

National Taiwan University, 1 Sec. 4, Roosevelt Rd., Taipei, 106319, Taiwan

Abstract

Deep learning models have been proposed to identify skin lesions based on the skin surface images. However, one unresolved issue is their lack of interpretability, which necessitates the development of skin lesion classification models capable of explaining the diagnostic features. Cell nuclei information plays a crucial role in skin lesion classification because it provides valuable insights into the underlying cellular changes associated with various skin diseases, aiding in accurate diagnosis and appropriate management of patients. This paper aims to identify quantitative features with cross-sectional cellular-resolution images to facilitate an interpretable model. We develop a melanin localization model that can be utilized on skin lesions. This model combines skin layers and cell nuclei segmentation models to derive a set of quantitative features. A multitask learning strategy is applied to enhance the segmentation accuracy and benefit from the shared information of these features. Subsequently, a tree-based machine learning model is employed to develop an interpretable classification model using these features. Using eczema as an example, optical coherence tomography that captures cellular structures, including nuclei *in vivo*, could potentially enhance understanding of pathogenesis and treatment response without requiring invasive biopsies.

Keywords

Multitask learning, skin lesions, optical coherence tomography, tree-based interpretable classification model

1. Introduction

Cell nuclei imaging is crucial for skin lesion classification for several essential reasons related to understanding cellular morphology and identifying abnormalities associated with various skin conditions, including cancers [1–3]. Several machine learning algorithms have been used for skin lesion classification, leveraging image analysis and pattern recognition to diagnose accurately [4–6]. These algorithms are typically trained on large datasets of skin lesion images annotated with corresponding diagnoses in the enfaces [7,8]. Physicians prefer using understandable algorithms in real-world scenarios, even though they usually have moderate, sometimes limited, performance compared to other complex black-box techniques [9]. In this work, the cross-sectional images of skin lesions were obtained using a cellular resolution optical coherence tomography (OCT) [10–14], which presents the opportunity for non-invasive and high-speed diagnosis and provides histopathological-level information. So far, minimal research has explored the human skin structures and lesions in OCT images. Several key quantitative features in the skin's cross-sectional planes can be readily acquired, e.g., nuclei, stratum corneum (SC), dermal-epidermal junction (DEJ), and melanin cluster. These features have shared information about the skin lesion characteristics. Multitask learning (MTL) aims to learn a shared representation that benefits all learning tasks [15–17]. By leveraging shared information across these features, we segment these features using MTL to improve the overall performance of each task.

The advancement of deep learning paves the way for systematically and automatically analyzing medical images. Therefore, this work aims to explore human skin cellular-resolution OCT imaging by developing a series of deep-learning approaches. Concurrently, explainable artificial intelligence (XAI) is now well-adapted to explain deep learning output on medical datasets [18,19].

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

tomohiroliu22@gmail.com (A. 1); shuang@ntu.edu.tw (A. 2)

0000-0003-0015-8652 (A. 1); 0000-0001-6244-1555 (A. 2)

Despite the availability of heat map visualizations, a comprehensive understanding of the model's decision-making process and the specific indicators it employs to identify skin lesions remains to be elucidated. In this work, our objective is to develop an explainable skin lesions classification model capable of providing quantitative insights into distinguishing different types of skin lesions.

2. Multitask segmentation model

To obtain quantitative features, we used supervised U-Net segmentation modes on nuclei, skin layers, and melanin, and subsequently, a tree-based machine learning model was employed to develop an interpretable classification model using these features. The images were collected from a clinical trial approved by IRB, Mackay Memorial Hospital, Taipei, Taiwan (# 20MMHIS039).

2.1. Cell nuclei and skin layer segmentation

We have presented a robust segmentation model trained on a diverse healthy and unhealthy skin image dataset [20,21]. Our model demonstrates exceptional accuracy in segmenting lesion images with proliferative SC and highly variable DEJ, as well as effectively detecting the distribution of melanin. However, it is essential to acknowledge that our model exhibits limitations in predicting malignant tumors, primarily due to the distinctive characteristics of such lesions and the scarcity of available data for training. As a result, the development of an explainable AI model specifically tailored to address these types of lesions is deemed less prioritized within the scope of this study.

2.2. Melanin localization

Although the skin layer and cell nuclei segmentation model effectively extracts information about the SC, DEJ, and keratinocytes, the diagnosis of most skin lesions relies heavily on identifying melanin amount and distribution. However, determining the morphology and location of melanin is typically challenging. Furthermore, the prior U-Net model primarily focuses on healthy human skin tissue, making it difficult to segment images featuring proliferative SC and undulated DEJ. Hence, we aim to develop a robust segmentation model that can highlight the position of melanin and effectively segment lesion images that feature abnormal SC and DEJ. The subsequent section will detail the dataset, methodology, model training, results, and analysis.

We propose an all-in-one segmentation model that simultaneously segments the air gap, SC, epidermis (excluding SC), dermis, cell nuclei, and melanin. We adopt an MTL scheme to train the model, which involves six tasks: tasks 1 to 5 are the segmentation of different combinations of nuclei, SC, DEJ, and dataset partitions. Task 6 is for melanin localization. By simultaneously training the model on these six tasks, the model can learn from multiple sources of information and obtain a more generalized result, mitigating the overfitting issue due to the small dataset.

In the experiment setup, we employed the overall model architecture in a U-Net architecture with 5 down-sampling layers, deep sharing [22,23], and deep feature sharing. The level of deep sharing was set to 2, resulting in a model with a total of 8 outputs, namely cell nuclei prediction, half-sized cell nuclei prediction, air gap prediction, SC prediction, epidermis* (excluding SC) prediction, epidermis prediction, dermis prediction, and melanin prediction. For training, we utilized the Adam optimizer with an initial learning rate of 0.001. The training epoch was set to 50, with a learning rate decay of 0.1 after 20 epochs and a batch size of 8. In each epoch, the model was trained on all 6 tasks sequentially.

The dataset used to train the robust segmentation model consists of six subsets. The first subset contains healthy human skin images cropped and horizontally flipped, resulting in 1224 images of size 384 × 512 pixels. The second subset comprises 20 normal skin, 20 solar lentigo, 20 nevus, and 20 vitiligo images annotated for melanin distribution. These images were also cropped and horizontally

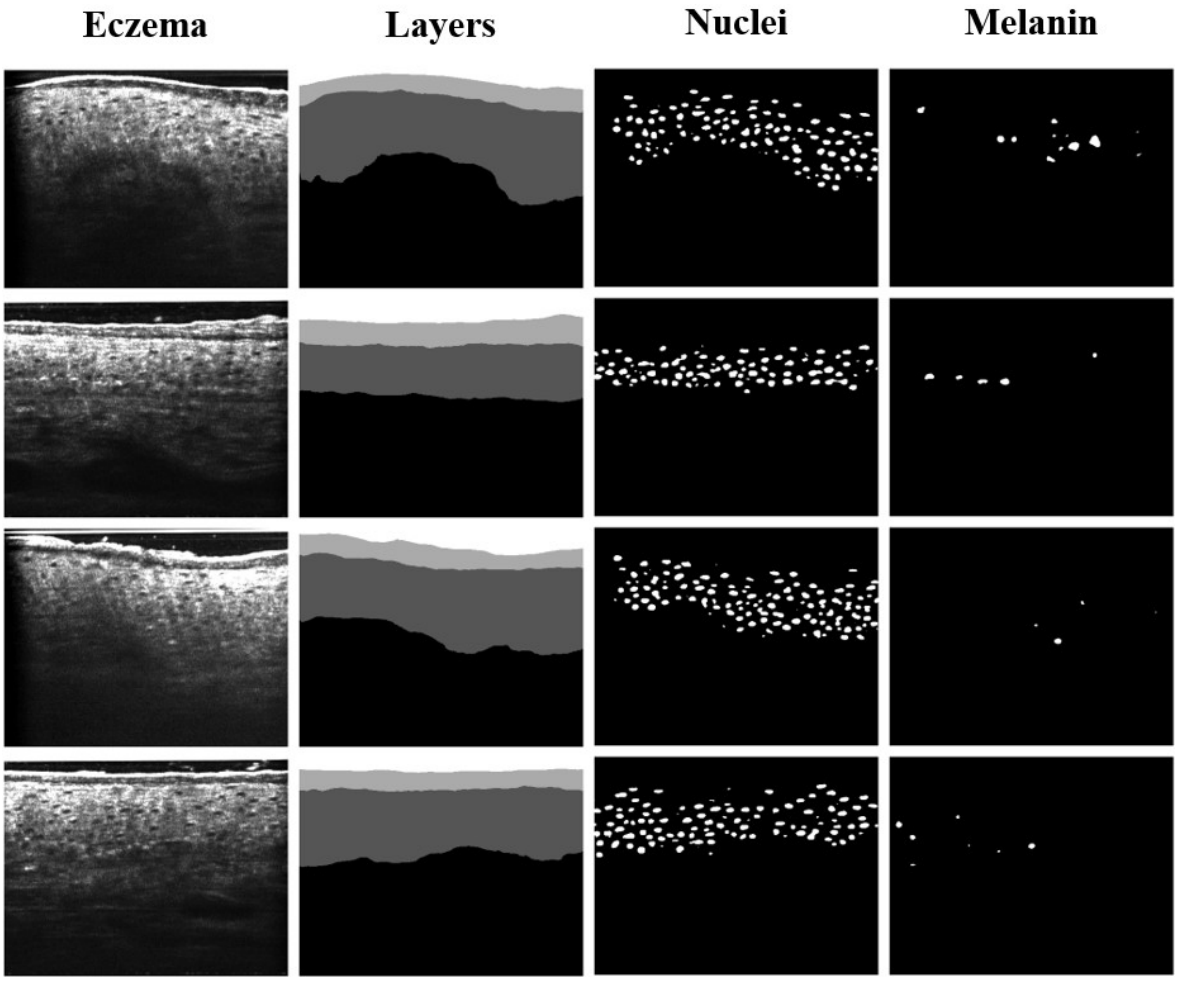

Figure 1: The segmentation results on eczema images using multitask learning. The first column demonstrates some examples of eczema OCT images. The second to fourth columns are predictions of the skin layer boundaries (skin surface, lower boundary of SC, and DEJ), nuclei, and melanin from the model trained with 6 tasks.

flipped, resulting in 320 images of size 384 × 512 pixels. The third subset includes 40 psoriasis and 40 eczema images annotated for SC labeling. These 80 images were converted into 320 images of size 384 × 512 pixels using the same image processing method. The fourth subset contains 40 solar lentigo and 40 seborrheic keratosis images annotated for SC and DEJ labeling. After cropping and horizontal flipping, these images were transformed into 320 images of size 384 × 512 pixels. The fifth and sixth subsets contain only normal skin images. The fifth subset includes 167 images annotated for cell nuclei, SC, and DEJ labeling, while the sixth subset includes 170 images annotated for cell nuclei and DEJ labeling. Figure 1 presents the segmentation results on eczema. We observed that the MTL model successfully segmented the SC, while the single task model inaccurately predicted a thin SC.

To generate a comprehensive set of features for training an explainable machine learning model, we conduct a statistical analysis to compare various parameters between normal and abnormal skin conditions. Specifically, the study includes measurements of SC thickness, epidermis thickness, cell nuclei area, and total melanin area. Due to the limited data availability for rare skin lesions, the statistical analysis primarily focuses on common skin abnormalities. The average SC thickness for normal skin is measured to be 11.97±3.83 μm, while for eczema, it increases to 14.31±8.25 μm. Notably, an observable increase in SC proliferation can be observed in cases of eczema characterized by its inflammatory nature. The mean epidermis thickness for normal skin measures 62.08±12.04 μm, while for eczema, it increases to 75.64±27.37 μm. This finding suggests that individuals with eczema exhibit

an increased epidermal layer thickness. The average cell nuclei areas of normal and eczema skins are 14.32±7.58 μm^2 and 15.44±8.43 μm^2, respectively.

3. Interpretable classification model

By segmenting different skin lesions based on morphological information about cell nuclei, skin layers, and melanin, these quantitative features can be fed into an explainable machine-learning algorithm to create a model that accurately identifies different skin conditions. Once the model has been trained, the results can be analyzed to gain insights into the performance, and the interpretability of the model allows us to understand the reasoning behind its predictions, which is particularly important in the medical field. By analyzing the results, we can identify areas where the model may need improvement and make necessary adjustments.

As a preliminary trial, we developed a model capable of accurately identifying different types of skin lesions by leveraging the morphological information obtained from OCT images. One challenge in interpreting CNN models is the high input data dimension. The images used for model training are cropped to a size of 384 × 1000 pixels, resulting in a feature dimension of 384,000 length. Analyzing each feature and interpreting the meaning of each pixel can be highly complicated. Moreover, the information in the image can be global or local, and it is not realistic to treat each pixel as an individual feature, as different pixels can be correlated. This correlation poses a significant obstacle in explaining the model's prediction and performance. Another difficulty in interpreting deep learning models is that the CNN architecture typically contains over a million parameters, making it difficult to understand the model's decision-making process.

To develop an interpretable model, it is necessary to individually design a series of quantitative features with low dimensions and select an explainable learning algorithm, such as XGBoost [24]. During the feature engineering stage, a total of 8 features were designed. Each cell nucleus in the segmentation result was collected for the cell nuclei feature, and the average area of the cell nuclei, the standard deviation of the cell nuclei area, and the average aspect ratio of the cell nuclei were calculated. The 3 features were designed to understand the influence of cell nuclei on lesion identification sincesome lesions have enlarged and lengthened cell nuclei in the upper epidermis.

For melanin, the summation of the area of the melanin in each image was computed. This feature is handy for validating that some skin lesions, such as nevus and solar lentigo, have an unusual amount of melanin. The thickness and undulation of the SC were calculated to provide quantitative value for evaluating the SC. Inflammatory diseases typically show a proliferative SC, so these SC-related features can be useful. Regarding the epidermis excluding SC classes, two related features were designed, namely the undulation and thickness of the epidermis. Seborrheic keratosis and psoriasis show an increased epidermis, and the average epidermis area and thickness can present the feature of these lesions. The standard deviation of the epidermis thickness was also computed to account for undulated DEJ in some lesions, such as solar lentigo.

Table 1 is organized and lists the description, notation, and calculation method for each feature to describe each feature more explicitly. We applied the XGBoost algorithm, which is designed to be scalable and efficient, even on large datasets. Moreover, the XGBoost automatically ranks the importance of features, facilitating the interpretation of results. Additionally, XGBoost supports parallel processing, which enables it to leverage GPUs for faster model training. The XGBoost framework is based on a gradient-boosting algorithm, gradually adding new models to improve the overall prediction. In each iteration, the algorithm calculates the gradient of the loss function relative to the current model's predictions and fits a new model to the residual errors. The new model is added to the ensemble, and the process continues until a stopping criterion is met. Our XGBoost incorporates advanced techniques, such as tree pruning and weighted quantile sketch, to improve traditional gradient boosting. Tree pruning removes parts of the tree that do not contribute to the overall performance, mitigating the risk of overfitting. A weighted quantile sketch is a data structure that efficiently computes the split points in decision trees, reducing computation time.

To better understand how our model makes predictions, we have employed the SHapley Additive exPlanations (SHAP) method [25]. The SHAP values are computed by comparing the model's predictions for a given instance with and without each feature. The difference between the predictions

Table 1
The feature engineering for 8 features used in the explainable lesion classification model.

Notation	Description	Calculation method
CELL_AREA_AVG	average area of the cellnuclei	collect all of the segmenting cell nuclei and calculate the average area
CELL_AREA_STD	standard deviation ofthe cell nuclei area	collect all of the segmenting cell nuclei and calculate the area standard deviation
CELL_ASEP_AVG	average aspect ratio ofthe cell nuclei	collect all of the segmenting cell nuclei and calculate the average value of width-to-length ratio
MELANIN	total amount of the melanin	summing all of the segmenting melanin area
SC_THICK	average thickness of theSC	calculating the difference between the highest and lowest point for SC along the horizontal direction and compute the average value
SC_UNDUL	undulation of the SC	calculate the length of the tissue surface curve along thehorizontal direction
EPI_THICK	average thickness of the epidermis (excluding SC)	calculating the difference between the highest and lowest point of the epidermis (excluding SC) along the horizontal direction and computing the average value
EPI_UNDUL	undulation of the epidermis (excluding SC)	calculate the length of the DEJ along the horizontaldirection

is then used to determine the contribution of each feature. The SHAP values are combined to explain the model's prediction for the instance. Our study used the SHAP method to produce a feature importance figure for each skin lesion case. These figures provide insights into the effect of each feature on the prediction. In addition, we have also included feature and SHAP value relation plots, which provide information on the distribution of data for each feature. By evaluating individual data with its corresponding SHAP value on all features, we can explain the decision-making behavior of our model. This approach allows for a more comprehensive understanding of how our model arrives at its predictions.

Using eczema as a demonstration, the model's overall accuracy on the training and testing sets is 96.38% and 78.57%, respectively. The model's accuracy in identifying normal skin and eczema is 74.63% and 90.16%, respectively. As shown in Fig. 2, the increased thickness of the epidermis is the most crucial factor in differentiating between eczema and normal skin images. The feature related to

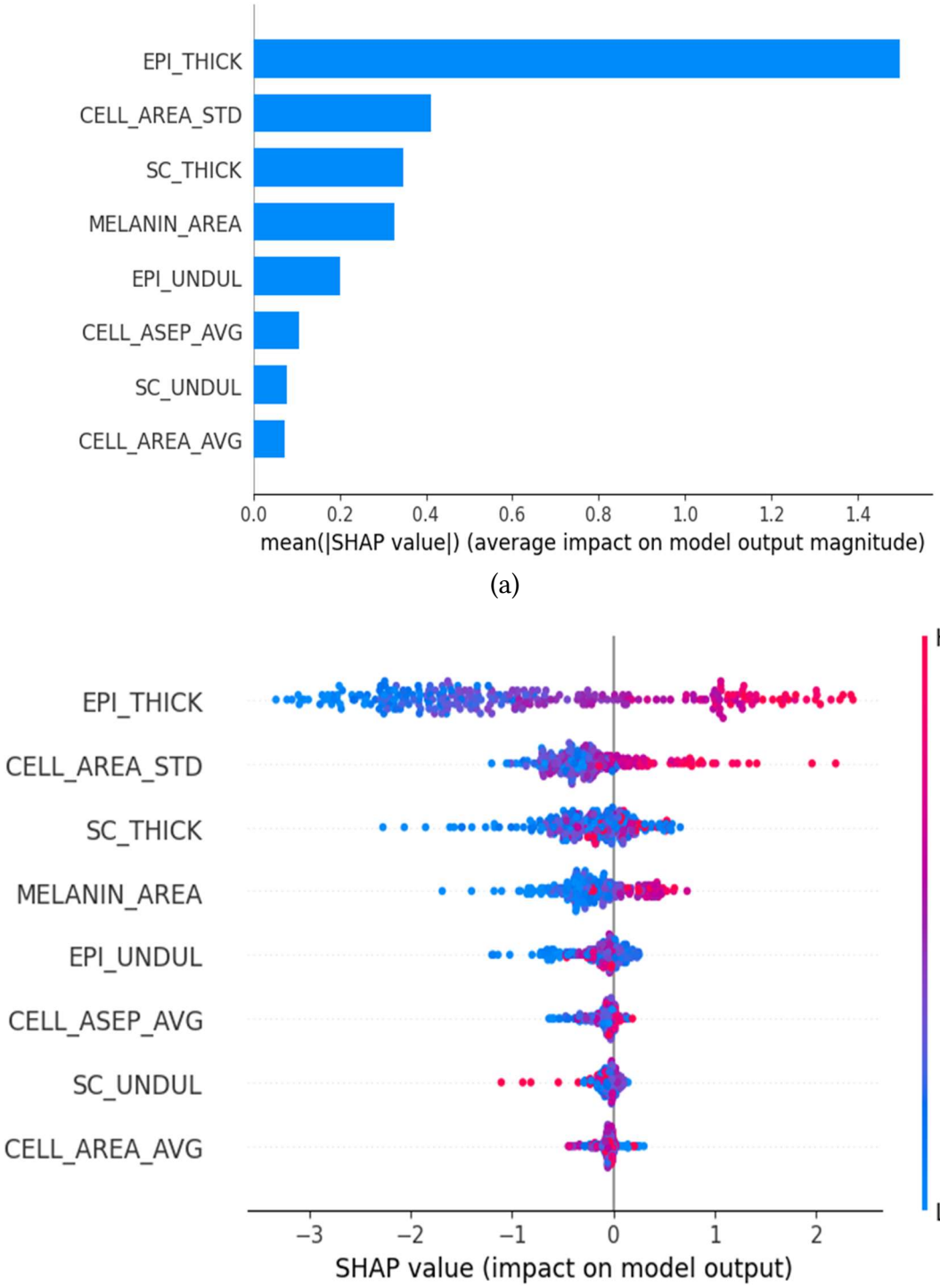

Figure 2: (a) The feature importance and (b) feature and SHAP value relation plot for eczema case.

epidermis thickness dominates the the model's decision-making process, as skin tissue with inflammatory disease tends to have a thicker epidermis. Additionally, the lengthened cell nuclei at the top of the epidermis layer and increased SC thickness are the second and third most important factors, respectively.

Moreover, we demonstrate the individual classification results in Fig. 3. The images in the first to third rows are correctly identified, while the image in the last column is incorrectly classified. The first, second, and third rows all show increased SC and epidermis; thus, their corresponding features

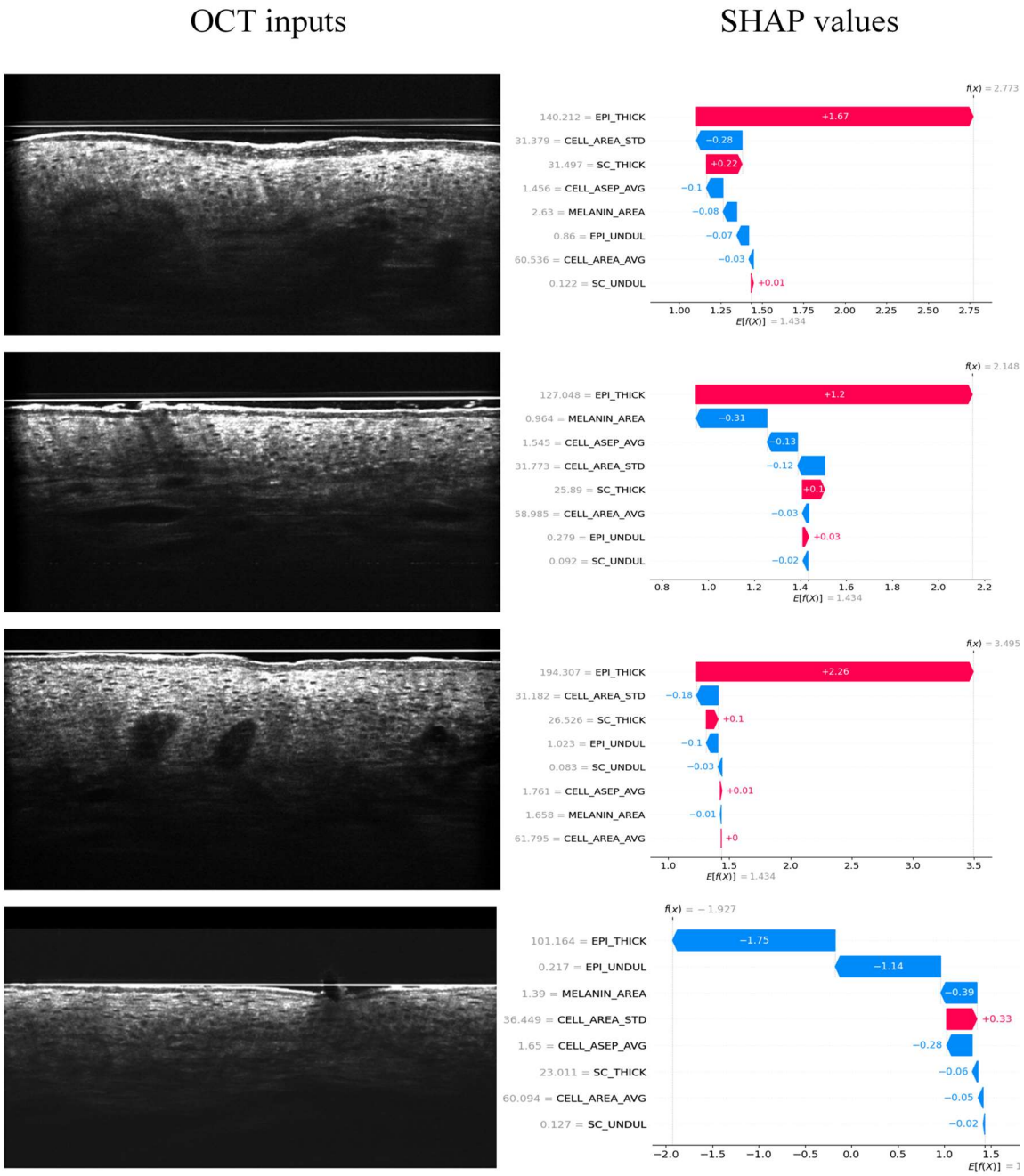

Figure 3: The eczema images with their corresponding SHAP values. The first to third rows are correctly classified, while the fourth row is incorrectly identified as normal skin.

appear as red bars. Since the eczema image in the last row does not show significant features, it is distinguished as normal skin by the XGBoost model.

We have tried the explainable lesion classification model based on the tree-based interpretable machine learning algorithm. We extract 8 quantitative features from the previous segmentation model predictions and evaluate our model on 6 different skin lesion categories, including eczema, psoriasis, solar lentigo, vitiligo, nevus, and seborrheic keratosis. The SHAP value explanation reveals that our experimental results agree with the deep learning model and medical knowledge. However, we also acknowledge the limitations of our feature selection and feature representation, which can significantly impact the model's accuracy.

4. Conclusion

Multitask learning is applied to improve the segmentation models for extracting crucial information about cell nuclei morphology, structural characteristics (SC and epidermis), and melanin content. These extracted features are then utilized to train an explainable machine learning model, providing quantitative insights into skin lesion analysis. The results obtained from the explainable skin lesion classification model demonstrate its ability to provide relative information for skin lesion diagnosis. While we have made progress in explainable clinical diagnosis, achieving trustworthy AI and precision medicine requires further advancements. A long-term project is needed to collect more diverse data and develop a comprehensive experimental design to enhance the capabilities of our models.

While cell nuclei imaging is not typically a routine diagnostic tool for eczema in clinical practice, it can be valuable for research purposes and in specialized settings where detailed cellular analysis is warranted. The primary diagnostic approach for eczema remains clinical assessment, often supplemented by laboratory tests to rule out other skin conditions or complications. The proposed explainable AI lesion identification model can help the physician understand the degree of deterioration for different skin and pave the way toward trustworthy and precise medical diagnosis using AI.

Acknowledgments

This work was supported by the National Science and Technology Council of Taiwan, grant numbers 110-2634-F-002-049, 110-2634-F-002-031, and 109-2634-F-002-025. The authors would also like to thank the National Center for High-performance Computing of National Applied Research Laboratories in Taiwan for providing computational and storage resources.

References

[1] D. S. Lidke, K. A. Lidke, Advances in high-resolution imaging – techniques for three-dimensional imaging of cellular structures, J. Cell Sci. 125 (2012) 2571–2580. doi.org/10.1242/jcs.090027.

[2] T. P. Medina, J. P. Kolb, G. Hüttmann *et al.*, Imaging inflammation – From whole body imaging to cellular resolution, Front. Immunol., 12 (2021). doi.org/10.3389/fimmu.2021.692222.

[3] M. E. McCarthy, M. R. Birtwistle, Highly multiplexed, quantitative tissue imaging at cellular resolution, Curr. Pathobiol. Rep. 7 (2019) 109–118. doi.org/10.1007/s40139-019-00203-8.

[4] P. Tschandl, N. Codella, B. N. Akay *et al.*, Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study, Lancet Oncol. 20 (2019) 938–47.

[5] N. Nigar, M. Umar, M. K. Shahzad, S. Islam and D. Abalo, A deep learning approach based on explainable artificial intelligence for skin lesion classification, IEEE Access 10 (2022) 113715–113725. doi: 10.1109/ACCESS.2022.3217217.

[6] A. Bhardwaj, P. P. Rege, Skin lesion classification using deep learning. In: Merchant, S.N., Warhade, K., Adhikari, D. (eds) Advances in Signal and Data Processing. Lecture Notes in Electrical Engineering, vol 703. Springer, Singapore. (2021). doi.org/10.1007/978-981-15-8391-9_42.

[7] B. Shetty, R. Fernandes, A. P. Rodrigues *et al.,* Skin lesion classification of dermoscopic images using machine learning and convolutional neural network, Sci. Rep. 12 (2022) 18134. doi.org/10.1038/s41598-022-22644-9

[8] N. Hameed, A. M. Shabut, M. K. Ghosh, M.A. Hossain, Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques, Expert Systems with Applications 141 (2020) 112961. doi.org/10.1016/j.eswa.2019.112961.

[9] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, Z. He, Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, Journal of the American Medical Informatics Association 27 (2020) 1173–1185. doi.org/10.1093/jamia/ocaa053.

[10] T. S. Ho, M. R. Tsai, C. W. Lu, H. S. Chang, S. L. Huang, Mirau-type full-field optical coherence tomography with switchable partially spatially coherent illumination modes, Biomed. Opt. Express 12 (2021) 2670–2683.
[11] C. Y. Tsai, C. H. Shih, H. S. Chu, Y. T. Hsieh, S. L. Huang, W. L. Chen, Sub-micron spatial resolution optical coherence tomography for visualising the 3D structures of cells cultivated in complex culture systems, Sci. Rep. 11 (2021) 3492.
[12] S. T. Tsai, C. H. Liu C. C. Chan, Y. H. Li, S. L. Huang, H. H. Chen, H&E-like staining of OCT images of human skin via generative adversarial network, Appl. Phys. Lett. 121 (2022) 134102.
[13] C. You, J. Y. Yi, T. W. Hsu, S. L. Huang, Integration of cellular-resolution optical coherence tomography and Raman spectroscopy for discrimination of skin cancer cells with machine learning, J. Biomed. Opt. 28 (2023) 096005.
[14] C. J. Ho, M. Calderon-Delgado, M. Y. Lin, J. W. Tjiu, S. L. Huang, H. H. Chen, Classification of squamous cell carcinoma from FF-OCT images: Data selection and progressive model construction, Computerized Medical Imaging and Graphics 93 (2021) 101992.
[15] R. Caruana, Multitask learning, Mach. Learn. 28 (1997) 41–75.
[16] F. Gao, H. Yoon, T. Wu, X. Chu, A feature transfer enabled multitask deep learning model on medical imaging, Expert Systems with Applications 143 (2020) 112957. doi.org/10.1016/j.eswa.2019.112957.
[17] Y. Zhang, Q. Yang, A survey on multitask learning, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 5586–5609. doi: 10.1109/TKDE.2021.3070203.
[18] S. Knapic, A. Malhi, R. Saluja, K. Framling, Explainable artificial intelligence for human decision support system in the medical domain, Mach. Learn. Knowl. Extract. 3 (2021) 740–770. doi: 10.3390/make3030037.
[19] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Medical Image Analysis 79 (2022) 102470. doi.org/10.1016/j.media.2022.102470.
[20] L. W. Fu, C. H. Liu, M. Jain, C. J. Chen, Y. H. Wu, S. L. Huang, H. H. Chen, Training with uncertain annotations for semantic segmentation of basal cell carcinoma from full-field OCT images, IEEE Transactions on Medical Imaging 43 (2024) 1060–1070.
[21] M. Jain, S. W. Chang, K. Singh, N. R. Kurtansky, S. L. Huang, H. H. Chen, C. S. J. Chen, High resolution full-field optical coherence tomography for the evaluation of freshly excised skin specimens during Mohs surgery: A feasibility study, J. Biophotonics (2023) e2023000275.
[22] H. Y. Chou, S. L. Huang, J. W. Tjiu, H. H. Chen, Dermal epidermal junction detection for full-field optical coherence tomography data of human skin by deep learning, Computerized Medical Imaging and Graphics 87 (2021) 101833.
[23] C. H. Liu, L. W. Fu, H. H. Chen, S. L. Huang, Toward cell nuclei precision between OCT and H&E images translation using signal-to-noise ratio cycle-consistency, Computer Methods and Programs in Biomedicine 242 (2023) 107824.
[24] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining (2016) 785–794.
[25] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, 30 (2017).