MULTI-LABEL PROTOTYPE BASED INTERPRETABLE MACHINE LEARNING FOR MELANOMA DETECTION

Afra Hussaindeen,
Department of Computer Science and Engineering,
University of Moratuwa,
Moratuwa, Sri Lanka
afrahussaindeen.17@cse.mrt.ac.lk

Shehana Iqbal
Department of Computer Science and Engineering,
University of Moratuwa,
Moratuwa, Sri Lanka
shehanaiqbal.17@cse.mrt.ac.lk

Thanuja D. Ambegoda
Department of Computer Science and Engineering,
University of Moratuwa,
Moratuwa, Sri Lanka
thanujaa@uom.lk

Submitted: *Mar, 05, 2022* ***Revised:*** *May,18, 2022* ***Accepted:*** *May, 24, 2022*

Abstract: Skin cancer is the most common of all cancers that exist in the world and melanoma is the deadliest among all the skin cancers. It is found that melanoma roughly kills a person every hour somewhere in the world. Considering the severity of the disease, significant effort goes into minimizing delays in the process of diagnosing melanoma. There are several approaches based on Machine Learning (ML) that can assist dermatologists in melanoma detection. However, many experts hesitate to trust ML systems due to their black-box nature, despite the accuracy of their performance. This highlights the need for applications that facilitate not only accurate classifications but also the ability to justify such decisions. In this work, we propose a prototype-based interpretable melanoma detector that uses the Seven Point Checklist, a well-known criterion used for the detection of melanoma. Prototypes provide the justification behind the decisions suggested by the ML model in a way of showing similar cases that are already known. In addition to identifying the dermoscopic features listed in the seven-point checklist, our work aims to provide reasoning that is similar to the ones used by the dermatologists in clinical practice for each decision made by the model. F1-Score has been used as the main performance metric in evaluating the model performance and that of the best performing class was 0.87. Furthermore, we show comparisons of our approach with Local Interpretable Model-Agnostic Explanations (LIME), a popular approach for interpretability for deep learning models.

Keywords: Skin lesion, dermatology, deep learning, interpretable ML, prototype-based, melanoma, seven-point checklist

I. INTRODUCTION

Melanoma is a type of skin cancer that occurs mainly in Melanocytes which cause the color of human skin. It can also take place in other parts of the body such as the eye, intestines, etc if they contain pigmentation issues. melanoma brings out visible signs such as the occurrence of new moles and the change of appearance of existing moles. Despite the seriousness of this disease, melanoma seems to be common among people in countries like the United States of America, New Zealand which proves that the disease is prevalent among white-skinned people. Unfortunately, about one out of every fifty Americans get diagnosed with melanoma in their lifetime [1].

Dermatology is a field of medicine that involves studying and specializing in the skin which is the largest organ of the human body, nails, and hair, and the medical conditions related to them. In order for dermatologists to save more lives of melanoma victims, it is extremely important to identify the disease at the early stages and direct the diseased to medications as soon as possible. Seven-point criterion, ABCD rule based on the criteria asymmetry (A), border (B), color (C), diameter (D), and pattern analysis are some of the widely accepted criteria among the practitioners and they are used initially to determine whether a skin lesion is benign or not before any further examinations. Seven Point Checklist as shown in Table 1 is a scoring algorithm that considers seven of the major contributing dermoscopic features to melanoma [2].

TABLE. 1 Seven Point Checklist

Feature	Score
Atypical pigment network	2
Blue-whitish veil	2
Atypical Vascular Pattern	2
Irregular pigmentation	1
Irregular streaks	1
Irregular dots and globules	1
Regression structures	1

Skin lesion image datasets that are available publicly for research related to dermatology mostly tend to contain either clinical images [3-6] or dermoscopic images [7-9]. Some of the datasets contain both of these types as well [10]. Derm7pt [10] is one of such publicly available datasets for research related to dermatology. It contains dermoscopic and clinical images of skin lesions along with metadata containing seven-point criterion-related information.

Many machine learning-based image classifiers have been introduced in several researches for skin image classification and especially melanoma detection throughout the past few years. These research works have focused on various concepts such as disease classification, pattern identification, concept-based prediction etc. Despite the high performance of classification, these existing machine learning-based skin lesion diagnosis systems still are not considered to be reliable. This is mainly due to its black-box nature which induces the inability of the systems to provide a human understandable reason behind its predictions. Interpretability is a crucial component of machine learning approaches that are introduced to serve the medical domain to address several problems in the field. There are various techniques like visualization, prototyping, feature statistics, etc. However, the amount of interpretability expected by a system will depend on the expertise on the subject of the end-users of the system to whom the reasoning will be provided by that model.

Deep learning, a class of machine learning methods, has been featured in many computer-aided diagnostic systems in medicine and healthcare. As a significant health issue in several countries, melanoma has been getting lots of attention from researchers, especially in terms of detecting it early to enable more effective treatment. As a result, there is a significant number of deep learning-based applications for machine learning based melanoma detection [11-13].

A computer-aided diagnosis system is proposed in [11] for melanoma detection using deep learning. The model has achieved its best accuracy of 99.1% on the PH2 dataset with the use of VGG-16 as the base architecture. The model has also been tested with the ISIC 2016 dataset and the model has performed better on both datasets after using augmentation techniques on them. Another deep learning-based approach is discussed in [12] that have overcome the limitations in the performance of typical applications due to the complexity of visual features in skin lesion images with the use of an encoder-decoder network. This network has been boosted with a multi-stage and multi-scale approach in order for the model to grab complex features and deal with different-sized skin lesions. It has performed with 95% accuracy on the ISIC 2017 dataset and 92% accuracy with the PH2 dataset.

Various researchers have attempted to address this issue by examining various parts of the topic. Some approaches have taken key conceptual features and criteria considered by dermatologists in identifying Melanoma [14-17]. Another set of researchers has used metadata like age, medical history, etc in addition to medical images to achieve better performance ([15, 18]). Skin lesion image classification with the use of Convolutional Neural Networks and disease-wise labeled skin images is also a common technique among most of the research works [19-20]. Some have also used hybrid approaches derived from the approaches mentioned earlier [15, 17]. An approach discussed in [14] for detecting skin cancer using the ABCD rule achieving a best accuracy of 84%. The CNN model in [15] outputs all the labels related to each category of features present in the Seven-Point Checklist, one of the heavily accepted criteria for identifying melanoma among practitioners. It has not only used dermoscopic images, but also clinical images and metadata to build the model which has achieved an AUROC of 89.6% at its best case. Another interesting work in [17] that has used both handcrafted and pre-trained CNN features to identify melanoma. Feature selection with genetic algorithms has been applied to improve the performance of the model and it has achieved its best performance of 98% overall accuracy on the PH2 dataset with the use of ResNet as the base architecture.

Interpretability is one of the crucial requirements when it comes to decision support systems in the medical domain due to the potentially severe outcomes of a misclassification and the need for the ML systems to be trusted by medical practitioners. Hence, ML researchers have begun to pay more attention to the interpretability of the machine learning systems which are to be used in decision making. Such approaches have achieved interpretability in different ways. Some have used concept-based techniques [21]. Interpretability via visualization techniques [22-24] such as heatmaps is another method that has been used in research in this domain. A novel CNN based pipeline is describe in [23] for Melanoma identification with the WSIs (Whole-slide images) along with a heat map-based interpretability approach. The Grad-CAM (Gradient-weighted Class Activation Mapping) method has been used to produce visual explanations for the classifications made by the model. The model has achieved an AUROC of 0.962 according to the evaluation results. Another recent work is discussed in [24] that has followed an ensemble approach for the diagnosis of Melanoma in skin lesions. This approach also has used heat maps as visual explanations for their classifications. Among the different ensembled models that have been tried out by the researchers, the best performing model has achieved an accuracy of 0.92.

Prototype-based interpretability is another technique that could bring user-friendly and reliable evidence, yet has not been used in the applications in the domain of dermatology so far. ProtoPNet [25] is a prototype-based interpretable deep learning-based multi-class image classifier that was originally developed for bird classification. The ProtoPNetis improved in [26] for medical imaging by introducing fine annotation loss and replacing max-pooling in ProtPNet by top-k average pooling to reduce each similarity map to a single similarity score. This approach has been applied to classify Mass Lesions in Digital Mammography and achieved an overall accuracy of 83%. In this work, we have made the following contributions:

- Developing an interpretable deep learning-based multi-label classifier to identify the key features mentioned in the seven-point checklist for a given skin lesion.
- Extending the ProtoPNet model which has originally been made for multi-class classification to support multi-label classification along with an improved version of the loss function.
- Providing human-understandable prototype-based evidence for the identification of dermoscopic features.

To the best of our knowledge, this is the first attempt to bring out a combination of the seven-point checklist, multi-label classification and prototype-based interpretability for diagnostic assistance in dermatology. The prototype-based explanation expected by our model for a dermoscopic image of a skin lesion will be as given in Figure 1.

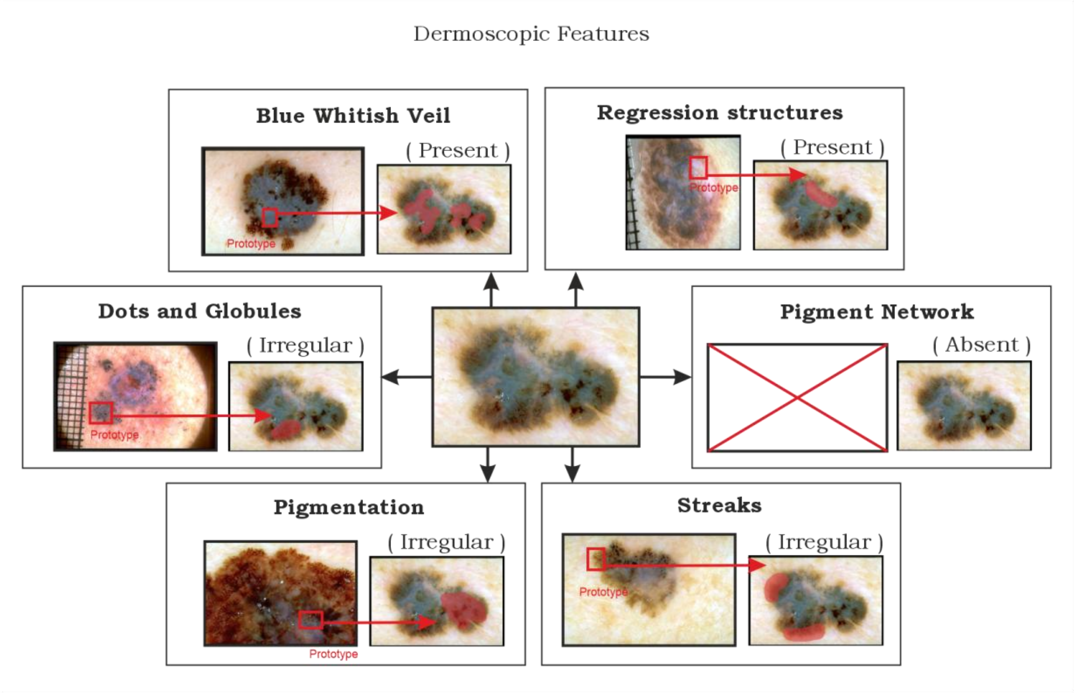

Fig. 1 Prototype based explanation expected by our model for a sample classification made by it for a melanoma Positive Skin lesion

II. METHODS AND MATERIALS

A. Dataset and preprocessing

The Derm7pt [10] dataset which comprises image data along with the metadata related to 1011 skin lesions was used for this work. We have built our dataset by extracting the dermoscopic images provided in the Derm7pt and classifying it mainly into seven classes and then into subclasses under each category. It is based on the seven-point criterion, also known as the seven- point checklist, a criterion introduced by Glasgow to help non-dermatologists identify dermoscopic features that contribute to possible melanoma. Each of the features has been assigned a score and the total score based on the presence or absence of these features will be used to determine if a skin lesion is malignant or benign.

Before being used for further steps, Image data was resized into 224 × 224 images and normalized and then categorized as given below. The data was split as the training set and test set in the 70:30 ratio for training and testing purposes. Table2 shows how we have formatted the dataset to align with our requirements.

TABLE. 2 Dataset Summary

Dermoscopic feature	**Classes (Derm7pt)**	**Classes (Our dataset)**
Pigment Network (PN)	Absent (400), Typical (381), Atypical (230)	Absent (400), Typical (381), Atypical (230)
Blue Whitish Veil (BWV)	Absent (816), Present (195)	Absent (816), Present (195)
Vascular Structures (VS)	Absent (823), Arborizing (31), Comma (23), Hairpin (15), Within regression (46), Wreath (2), Dotted (53), Linear regular (18)	-
Pigmentation (PIG)	Absent (588), Diffuse regular (115), Localized regular (3), Diffuse irregular (265), Localized irregular (40)	Absent (588), Regular (118), Irregular (305)
Streaks (STR)	Absent (653), Regular (107), Irregular (251)	Absent (653), Regular (107), Irregular (251)
Dots and Globules (DaG)	Absent (653), Regular (107), Irregular (251)	Absent (653), Regular (107), Irregular (251)
Regression Structures (RS)	Absent (758), Blue areas (116), White areas (38), Combinations (99)	Absent (758), Present (253)

In our research due to lack of sufficient samples, we excluded the vascular structures, and combined a few other categories into one.

B. Proposed Approach

We have extended the ProtoPNet [25] architecture to extract dermoscopic features. The main improvements are multi-label classification support and use of topmost-k average pooling instead of max pooling when computing the similarity score. The use of topmost-k average pooling was inspired by IAIA-BL [26].

As shown in Figure 2, the proposed interpretable model is composed of a sequence of convolutional layers$f$, prototype layers p, and fully connected layers h. The expected input to the model at the test time is a skin lesion image. During training, apart from skin lesion image, a multi-label label-encoded feature vector which consists of 6 main criteria given in the seven-point checklist each indicating the class index of either absence, presence, or the sub-type of the dermoscopic feature is also provided. The convolutional layers f used in our model are from the ResNet-152 [27] network pre-trained on ImageNet [28], and it extracts meaningful convolutional featuresto perform classification from a dermoscopic image.

Following the convolutional layers f, the prototype layer g contains m prototypes, $P = (P_j)_{j=1}^{m}$ learned from the training set. P stands for the whole set of prototypes learned for each dermoscopic feature while P_j represents a prototype of a dermoscopic feature. For an example if we enforced the model to learn 3 prototypes per each dermoscopic feature, then P_0, P_1, P_2 will represent the prototypes for the first dermoscopic feature,P_3, P_4, P_5 will represent the prototypes for the second considered dermoscopic feature and so on. Here it is to be noted that each prototype represents a pattern of prototypical activation in a convolutional output feature maps' patch, which will then correlate to a patch of an original image. As a result, each learned prototype P_j can be considered as a representation of a unique dermoscopic feature. The model intends to learn prototypes only for the presence of dermoscopic features leaving the absence cases to be handled at the last layer h that is composed of multiple fully connected layers corresponding to each criterion in the seven-point checklist.

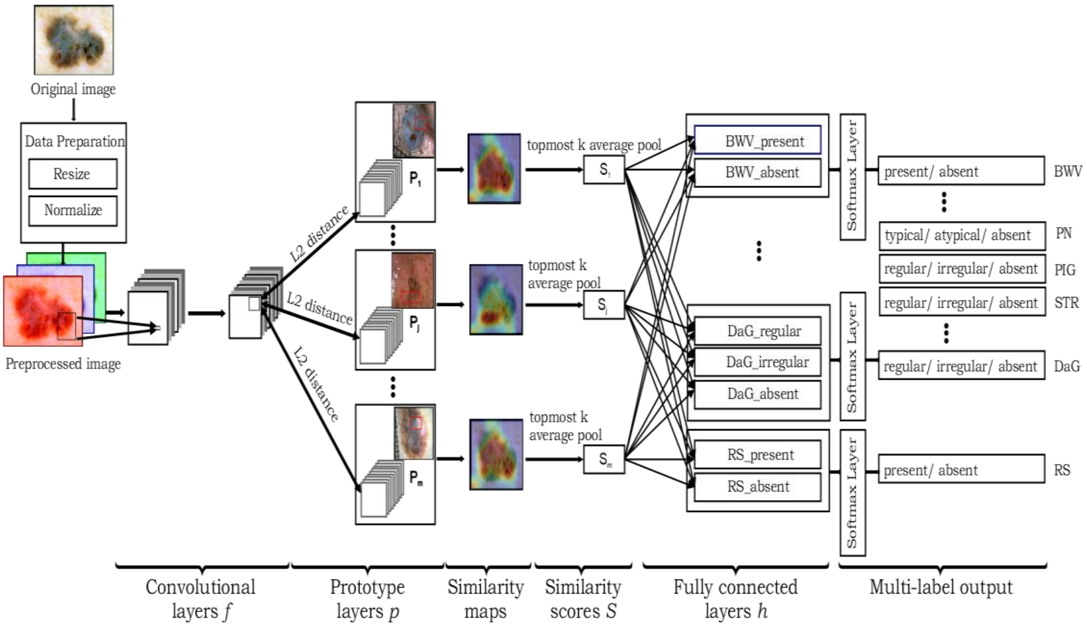

Fig. 2 Overview of proposed interpretable model architecture

For a given image x, the model first extracts the convolutional feature maps$f(x)$and then using the prototype layer$g$it computes a patch specific similarity distance $d_{j,i}$considering the squared Euclidean distances between each 1×1 patches of convolutional feature maps $f(x)$and each learned prototype P_j using Eq. 1 same as used in [26]. Next using Eq. 2 same as used in [26], the patch specific similarity distance $d_{j,i}$are converted into a patch specific similarity score $s_{j,i}$.

$$d_{j,i} = \left\| z - P_j \right\|_2^2 \tag{1}$$

$$s_{j,i} = \log \frac{d_{j,i}+1}{d_{j,i}+\varepsilon} \tag{2}$$

where i is the index of the 1×1patches of the14×14convolutional feature maps $f(x)$and z is the i^{th} 1x1 patch of convolutional feature maps $f(x)$.

The outcome of the prototype layer g is a similarity activation map which shows how prominent a prototypical part is in the image. This activation map retains the spatial relationship of the convolutional outputs and is later upsampled to the input image's scale to provide an overlaid similarity map that displays which part of the input image is most aligned with the learned prototype. Unlike ProtoPNet [25], our model takes into account the top 10% (i.e., $k = \frac{10}{100} \times (14 \times 14) \sim 19$) of the most activated convolutional patches that are closest to each prototype, rather than just the topmost activated patch as in ProtoPNet [25]. Using topmost-k average pooling inspired by IAIA-BL [26] as in Eq. 3, the activation map of patch-specific similarity scores obtained for prototype P_j is then condensed to an image-specific similarity score S_j. The topmost-k average pooling represented by $averagePool\,(topmost_k(\,[s_{j,i}]_{i=1}^{14\times14}, k\,))$is computed by taking into account the average of highest k patch-specific similarity scores. This indicates how prominent a particular feature can be widely spread in the skin lesion, which is a quite common occurrence for some of the features that we have considered.

$$S_j = averagePool\,(topmost\_k(\,[s_{j,i}]_{i=1}^{14\times14}, k\,)) \tag{3}$$

Following the computation of all image-specific similarity scores S_j for each prototype P_j , the last layer h composed of multiple fully connected layers followed by softmax layers were utilized to determine the probability of presence or absence of each sub-type of dermoscopic features. Finally, for each criterion, the class (i.e., either absent class or a sub-type of the dermoscopic feature) which has the highest probability is considered as an exact match for that criterion. The interpretability on the identification of dermoscopic features is then provided using the overlaid similarity map, together with the prototype.

During the training stage, the model seeks to learn a meaningful latent space, where the most significant patches for identifying dermoscopic features are clustered around semantically similar prototypes of the relevant classes while ensuring a good separation between clusters of different classes. In our research, the optimization problem is defined slightly differently than in [25] & [26], due to the fact that we support multi-label classification. The modified optimization problem is,

$$\theta, P\ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_l}\sum_{l=1}^{n_l} CrossEntropy\,(h_l\,.\,averagePool\,.topmost\_k.\ g\,.\,f(x_i)\,, y_i[l]) + \lambda_1 ClusterCost + \lambda_2 SeparationCost \tag{4}$$

$$ClusterCost = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_l}\sum_{l=1}^{n_l} \min_{j:class\,(P_j)=y_i[l]}(\frac{1}{k}\sum mink_{z\,\in\,patc\,hes(f(x_i))}\,(\left\|z - P_j\right\|_2^2)) \tag{5}$$

$$SeparationCost = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_l}\sum_{l=1}^{n_l} \min_{j:class\,(P_j)\neq y_i[l]}(\frac{1}{k}\sum mink_{z\,\in\,patc\,hes(f(x_i))}\,(\left\|z - P_j\right\|_2^2)) \tag{6}$$

where θ represents the parameters of the convolutional layers, n is the number of images, n_l is the number of labels, in our case it is the 6 criterias in the seven-point checklist, $y_i[l]$ is the index of the sub-type dermoscopic feature specific to l^{th} criteria in the seven-point checklist for the i^{th}image, $(\frac{1}{k}\sum mink_{z\,\in\,patc\,hes(f(x_i))}\,(\|z - P_j\|_2^2)$ gives the average of the minimum k patch specific similarity distances, λ_1 and λ_2 are constants. Differing from [25] & [26] in our work we implemented the cluster cost Eq. 5 such that the model learns at least one prototype for each sub-type dermoscopic feature that comes under each criterion specified in the seven-point checklist from the training image dataset.

III. RESULTS AND DISCUSSIONS

In this work, our main focus is to evaluate the applicability of the prototype-based interpretability approach to the field of dermatology. Hence, the model's performance was assessed quantitatively as well as qualitatively. Due to the presence of class imbalance in the dataset, we used the evaluation metric F1-score as in Eq. 9 to quantify the model's performance and for the qualitative evaluation, we relied on the dermatologist's expertise to validate the learned prototypes.

$$Precision = \frac{True\ Positives}{True\ Positives\ + False\ Positives} \tag{7}$$

$$Recall = \frac{True\ Positives}{True\ Positives\ + False\ Negatives} \tag{8}$$

$$F1\,Score = 2 \times \frac{Precision\ \times Recall}{Precision\ + Recall} \tag{9}$$

At the time of training, we enforced the model to learn three prototypes for each subcategory of each dermoscopic feature listed out in the seven-point checklist. Figure 3 depicts the overall performance of our model in terms of quantitative analysis while Figure 4 and Figure 5 show few of the learned prototypes classified based on the dermatologist's verification. According to dermatologist verification, 27% of the overall learned prototypes are accurate, while 73% of the prototypes are either from healthy skin regions or incorrect prototypes learned from the skin lesion region. Further, 27% of all incorrect prototypes come from healthy skin areas. This emphasizes the necessity of enforcing the model to focus solely on the skin lesion region to enhance the model's performance in learning accurate prototypes.

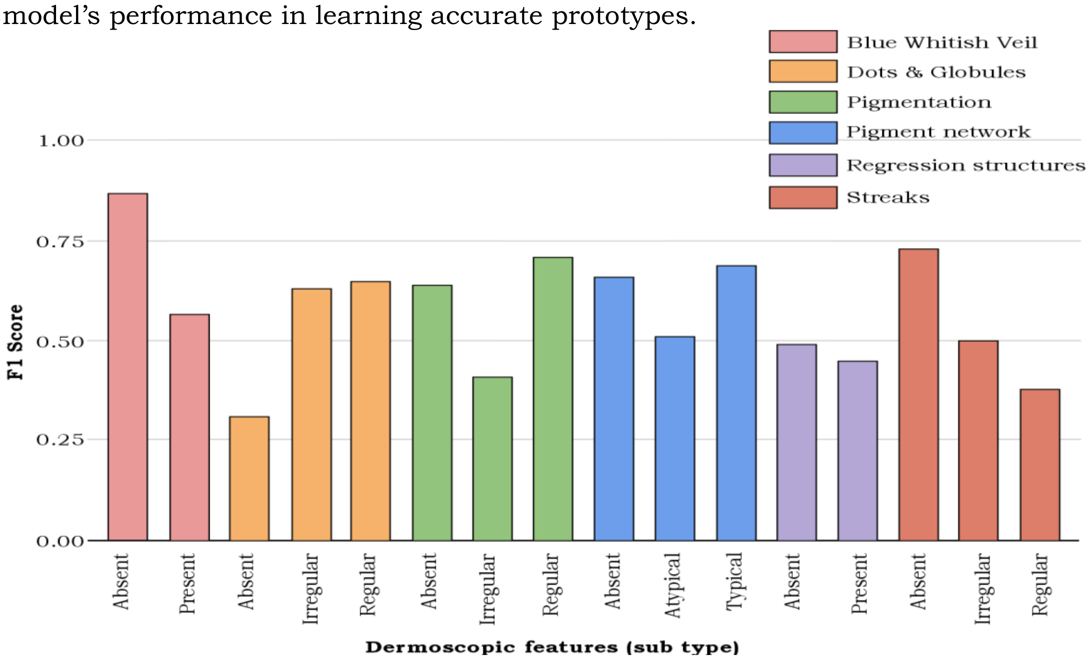

Fig. 3 Overall performance of the proposed model

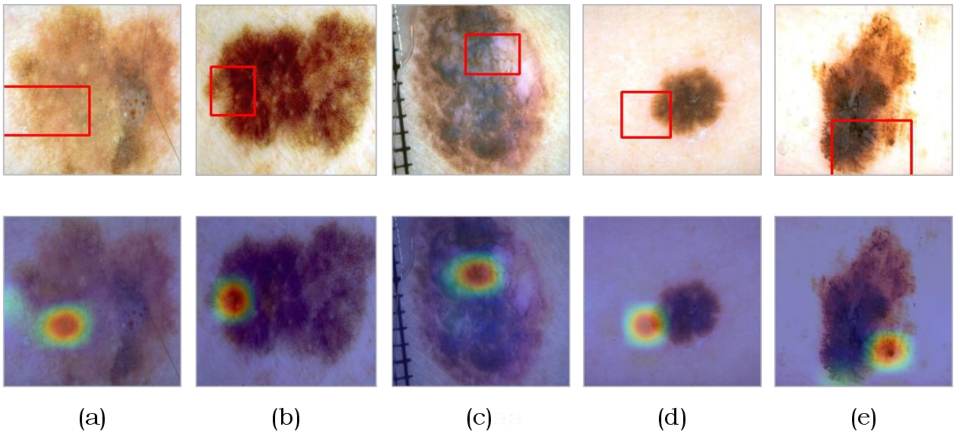

Fig. 4 Few learned accurate prototypes in the original image are represented by the top line, while the corresponding prototypical activation is represented in the bottom line. (a) Typical pigment network (b) Atypical pigment network (c) Presence of regression structures (d) Regular streaks (e) Irregular streaks

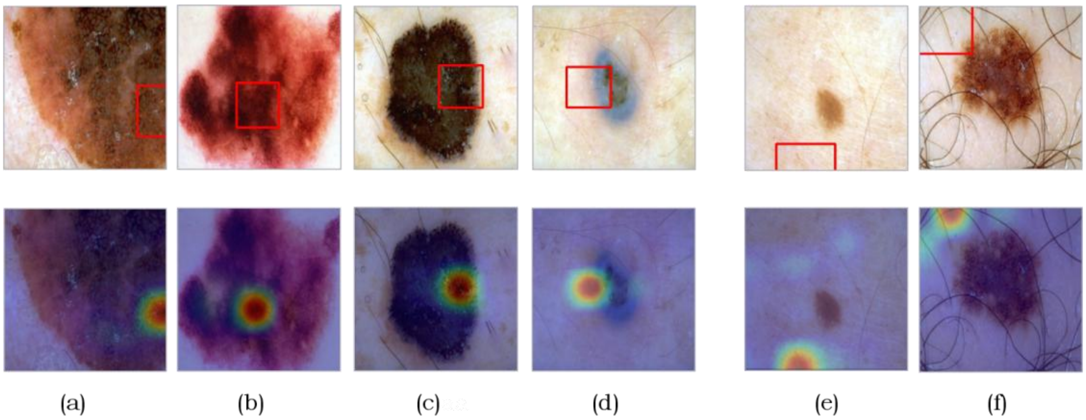

Fig. 5 Few learned inaccurate prototypes in the original image are represented by the top line, while the corresponding prototypical activation is represented in the bottom line. The inaccurate prototypes are from both skin lesion region as well as from healthy skin region and according to our model, (a) Irregular dots and globules (b) Atypical pigment network (c) Irregular streaks (d) Blue whitish veil (e) Regular pigmentation (f) Regular dots and globules.

To demonstrate how our model performs reasoning, we fed the model a dermoscopic image of a melanoma skin lesion. Figure 6 depicts the model's classifications for that melanoma skin lesion, as well as evidence for each classification. As previously stated, here the evidence is provided in the form of a heatmap overlayed on the original image, indicating the region that is highly similar to the learned prototype. Consider the first row, which represents the blue whitish veil. Because the region highlighted in red in the fed melanoma skin lesion looks almost identical to the region bounded by the red box in the learnt prototype for blue whitish veil, the model concludes that the blue whitish veil is present in the skin lesion. The same strategy was used to justify the identification of the remaining features as well.

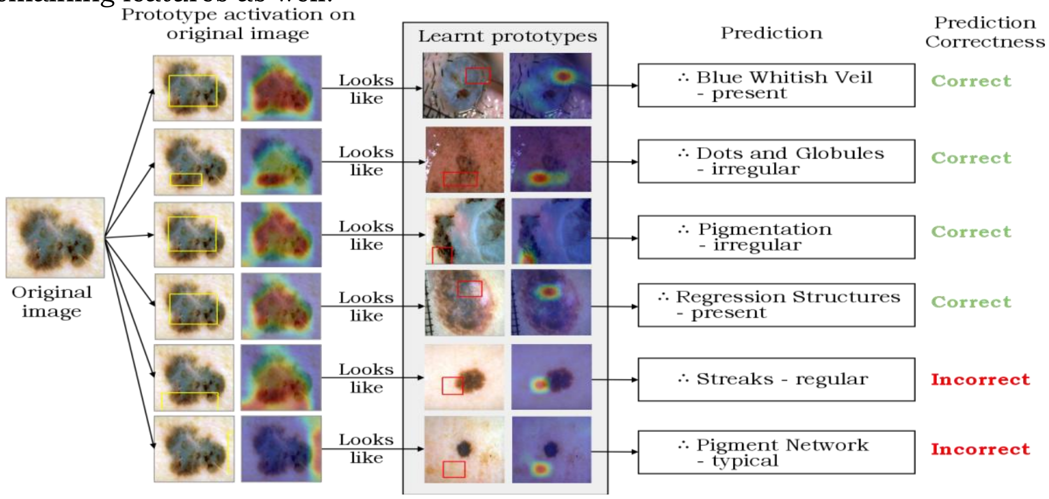

Fig. 6. Classification explanation generated by our model for a melanoma skin lesion test image

In order to compare our approach with the state-of-the-art approach for interpretable machine learning, we used the LIME [29] library on a deep learning model that uses ResNet152 [27] to identify the presence of the Blue Whitish Veil dermoscopic feature which is present in the seven-point checklist in skin lesions. The Derm7pt [10] dataset was used for this model. The model performed with an overall accuracy of 56.71%. Compared to the F1-score of our work for the Blue Whitish Veil feature which is 0.87 and 0.57 for the absent and present classes respectively, our prototype-based model has outperformed the typical deep learning classifier we used for the LIME [29] based approach by 0.18 and 0.31 for the absent and present classes respectively. Although the model outperforms the classical classifier in this case, there are deep learning-based classifiers built for melanoma identification that perform better than our interpretable model presented by previous research works as well [11-12]. Figure 7 shows a classification explanation provided by each model for a melanoma skin lesion. Here the explanations provided by a LIME [29] based image classifier can be visualized as an overlay along with the image samples or in several other similar ways

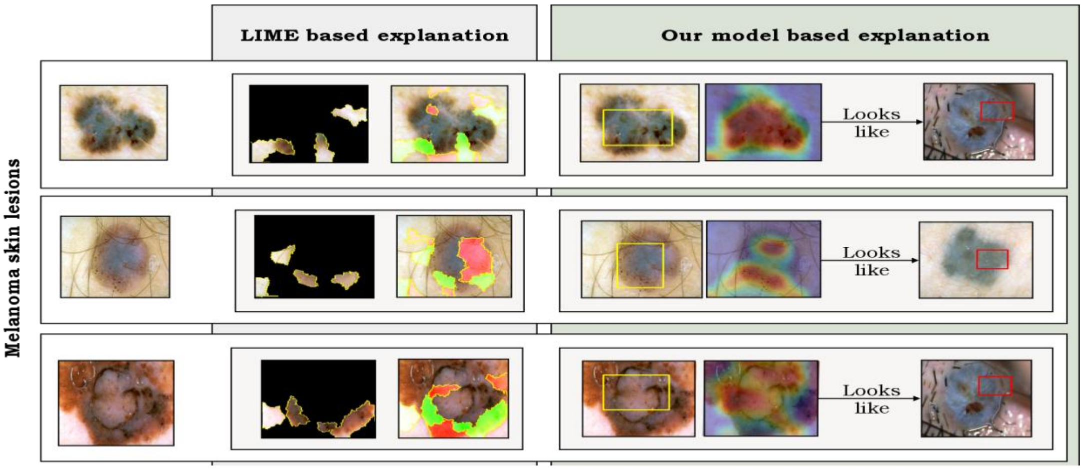

Fig. 7. Prediction explanation generated from a LIME based model and our model for a melanoma skin lesion. The explanation from our model which uses a learnt prototype for explanation is more descriptive compared to LIME.

Green color masks in Figure 7 show that the parts of the image that might have contributed to the classified class label and the parts of the image masked in red might have contributed to the classification negatively. Unlike our prototype-based approach, the LIME [29] based model that we built for the detection of Blue Whitish Veil dermoscopic feature's presence only highlighted the areas of the skin lesion where the feature might be available and where it might not be available without any prototypes to support the dermatologist analyze the correctness of the classification.

IV. CONCLUSION

Interpretable outputs of a machine learning model are of utmost importance for medical diagnostics because it allows the medical practitioners to understand the reasoning behind the decision produced by such a diagnostic support tool. This work aims to introduce prototype-based interpretability for machine learning based diagnostic assistance for melanoma, which is one of the deadliest skin diseases that prevail among people worldwide. Here, a prototype refers to an example derived from the training data that looks similar to what the model tries to predict. During the comparison with LIME, another popular approach to achieve interpretability, we observed that our model provides more descriptive and meaningful explanations. In order to improve the performance of the model i.e., to get more informative prototypes more consistently, we believe a segmentation step before the prototype extraction step (which needs segmentation labels to learn the boundary of the lesion) would be helpful.

Acknowledgements: We would like to thank Dr. Achala Liyanage, Senior Lecturer and Specialist Dermatologist, Department of Community Medicine, Faculty of Medicine, University of Ruhuna, Sri Lanka for participating in qualitatively analyzing our model performance.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1]. American Cancer Society, “Key statistics for melanoma skin cancer,” 2021.

[2]. Dermoscopedia, “Seven point checklist,” 2017.

[3]. X. Sun, J. Yang, M. Sun and K. Wang, “A benchmark for automatic visual classification of clinical skin disease images,” European Conference on Computer Vision, 2016, pp. 206–222.

[4]. A. G. Pacheco, G. R. Lima, A. S. Salomao, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro and F. B. Rodrigues, “PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones,” Data in brief, vol. 32, 2020, pp. 106221.

[5]. S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park and S. E. Chang, “Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm,” Journal of Investigative Dermatology, vol. 138, no. 7, 2018, pp. 1529–1538.

[6]. B. Xie, X. He, S. Zhao, Y. Li, J. Su, X. Zhao, Y. Kuang, Y. Wang and X. Chen, “XiangyaDerm: a clinical image dataset of Asian race for skin disease aided diagnosis,” Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention, 2019, pp. 22–31.

[7]. T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. Marcal and J. Rozeira, “PH^2- A dermoscopic image database for research and benchmarking,” 35th annual international conference of the IEEE engineering in medicine and biology society, 2013, pp. 5437–5440.

[8]. I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman and N. Petkov, “MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images,” Expert systems with applications, vol. 42, no. 19, 2015, pp. 6578–6585.

[9]. P. Tschandl, C. Rosendahl and H. Kittler, “The HAM10000 dataset, a large collection of multi-sourcedermatoscopic images of common pigmented skin lesions,” Scientific data, vol. 5, no. 1, 2018, pp. 1–9.

[10]. J. Kawahara, S. Daneshvar, G. Argenziano and G. Hamarneh, “Seven-point checklist and skin lesion classification using multitask multimodal neural nets,” IEEE journal of biomedical and health informatics, vol. 23, no. 2, 2018, pp. 538–546.

[11]. L. Singh, R. R. Janghel and S. P. Sahu, “A deep learning-based transfer learning framework for the early detection and classification of dermoscopic images of melanoma,” Biomedical and Pharmacology Journal, vol. 14, no. 3, 2021, pp. 1231–1247.

[12]. A. A. Adegun and S. Viriri, “Deep learning-based system for automatic melanoma detection,” IEEE Access, vol. 8, 2019, pp. 7160–7172.

[13]. J. A. A. Salido and C. Ruiz, “Using deep learning to detect melanoma in dermoscopy images,” International Journal of Machine Learning and Computing, vol. 8, no. 1, 2018, pp. 61–68.

[14]. E. M. Senan and M. E. Jadhav, “Analysis of dermoscopy images by using ABCD rule for early detection of skin cancer,” Global Transitions Proceedings, vol. 2, no. 1, 2021, pp. 1–7.

[15]. J. Kawahara, S. Daneshvar, G. Argenziano and G. Hamarneh, “Seven-point checklist and skin lesion classification using multitask multimodal neural nets,” IEEE journal of biomedical and health informatics, vol. 23, no. 2, 2018, pp. 538–546.

[16]. G. Di Leo, A. Paolillo, P. Sommella and G. Fabbrocini, “Automatic diagnosis of melanoma: a software system based on the 7-point check-list,” 43rd Hawaii international conference on system sciences, 2010, pp. 1–10.

[17]. X. Sun, J. Yang, M. Sun and K. Wang, “A benchmark for automatic visual classification of clinical skin disease images,” European Conference on Computer Vision, 2016, pp. 206–222.

[18]. X. Dai, I. Spasic´, B. Meyer, S. Chapman and F. Andres, “Machine learning on mobile: An on-device inference app for skin cancer detection,” Fourth International Conference on Fog and Mobile Edge Computing, 2019, pp. 301–305.

[19]. K. M. Hosny, M. A. Kassem and M. M. Fouad, “Classification of skin lesions into seven classes using transfer learning with AlexNet,” Journal of digital imaging, vol. 33, no. 5, 2020, pp. 1325–1334.

[20]. M. A. Kassem, K. M. Hosny and M. M. Fouad, “Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning,” IEEE Access, vol. 8, 2020, pp. 114822–114832.

[21]. A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed, “On interpretability of deep learning based skin lesion classifiers using concept activation vectors,” International Joint Conference on Neural Networks, 2020, pp. 1–10.

[22]. X. Li, J. Wu, E. Z. Chen and H. Jiang, “What evidence does deep learning model use to classify skin lesions?,” arXiv preprint arXiv:1811.01051, 2018.

[23]. P. Xie, K. Zuo, J. Liu, M. Chen, S. Zhao, W. Kang and F. Li, “Interpretable Diagnosis for Whole-Slide Melanoma Histology Images Using Convolutional Neural Network,” Journal of Healthcare Engineering, vol. 2021, 2021, pp. 1-7.

[24]. I. A. Alfi, M. M. Rahman, M. Shorfuzzaman and A. Nazir, “A Non-Invasive Interpretable Diagnosis of Melanoma Skin Cancer Using Deep Learning and Ensemble Stacking of Machine Learning Models,” Diagnostics, vol. 12, no. 3, 2022, pp. 1–18.

[25]. C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su and C. Rudin, “This looks like that: deep learning for interpretable image recognition,” arXiv preprint arXiv:1806.10574, 2018.

[26]. A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin, “IAIA-BL: A case-based interpretable deep learning model for classification of mass lesions in digital mammography,” arXiv preprint arXiv:2103.12308, 2021.

[27]. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[28]. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[29]. Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "" why should you trust my explanation?" understanding uncertainty in lime explanations," arXiv preprint arXiv:1904.12991, 2019, pp. 1-10.