



Research article

An interpretable hierarchical semantic convolutional neural network to diagnose melanoma in skin lesions

Hui-Ching Wu¹, Yu-Chen Tu², Po-Han Chen² and Ming-Hseng Tseng^{2,3,*}

¹ Department of Medical Sociology and Social Work, Chung Shan Medical University, Taichung 402, Taiwan

² Department of Medical Informatics, Chung Shan Medical University, Taichung 402, Taiwan

³ Information Technology Office, Chung Shan Medical University Hospital, Taichung 402, Taiwan

***Correspondence:** Email: mht@csmu.edu.tw; Tel: +886424730022 ext. 12214.

Abstract: Melanoma is a more dangerous skin cancer than other types of skin cancer because it rapidly spreads to other organs in its early stages. In the increasingly popular task of computer-aided diagnosis using deep learning methods, these models are difficult to interpret and often considered “black boxes”. The lack of interpretation of the model prevents the target users from fully understanding it. This study proposes a new interpretable hierarchical semantic convolutional neural network (MEL-HSNet) to diagnose melanoma. The benefits and strength of our approach are a white-box model that not only predicts whether a skin lesion observed in a dermoscopy scan image is melanoma but also provides explanatory information for decision-making. Compared to other convolutional neural networks, the MEL-HSNet model proposed in this study can generate interpretable information on melanoma prediction and obtain significantly better results compared to the other available models.

Keywords: melanoma classification; skin cancer; deep learning; convolutional neural networks; model interpretability; intelligent; healthcare

1. Introduction

Gradually, cancer has become one of the leading causes of death; it poses a significant barrier to increasing life expectancy. In many countries, cancer deaths have exceeded those from stroke or coronary heart disease [1]. According to the World Health Organization (WHO) statistical report

for 2019, cancer has become the leading cause of death before age 70 in 62% of countries around the world [2].

Melanoma, also known as malignant melanoma, is a skin cancer characterized by abnormal proliferation of melanocytes and invasion of other tissues. Malignant melanoma can develop from congenital or acquired benign melanocytic nevi, or malignantly from dysplastic nevi; however, most melanomas are de novo. Melanoma affects the skin, mucous membranes, and internal organs; it is more common in adults and is rarely reported in children. Recently, the incidence of malignant melanoma and the mortality rate have increased [3]. Compared to other types of cancers, the age pattern of mortality began to shift to younger age groups. Furthermore, delays in seeking medical care worsen the prognosis, leading to metastasis and even death [3,4].

Therefore, the early diagnosis and treatment of malignant melanoma are extremely important. Melanoma generally presents as a mole of the skin with asymmetric growth, irregular borders, color changes, a diameter of > 6 mm, and changes in appearance in a few weeks or months. These five diagnostic features are collectively called ABCDE [5]. A self-examination can be performed at home, preferably with the help of a partner, close friend, or family member following ABCDE criteria; however, this cannot replace the role of a dermatologist who performs a more thorough examination with a dermoscope. A dermoscope, known as a transdermal light microscope, is a non-invasive skin microscope. Some studies have demonstrated that dermoscopy has extremely high sensitivity and specificity for diagnosing acral melanoma and that it can make correct diagnoses in extremely early stages when the naked eye cannot [6,7]. In addition, it is a valuable tool for dermatologists when it comes to diagnosing early malignant melanoma. Therefore, in this study, an AI model was developed for a computer-aided diagnosis (CAD) system for malignant melanoma based on dermoscopic images.

Machine learning classifiers were used to develop automatic diagnosis methods for skin lesions [8,9]. Before modeling these classifiers, a set of image features must be manually segmented, such as skin lesion-related characteristics to which dermatologists pay special attention. However, recently, in most computer vision tasks, DL convolutional neural networks (CNN) can automatically extract high-level features and significantly improve classification performance. Therefore, a CAD system based on the CNN framework, which has been widely used by modern researchers, was used to detect multiple diseases [10,11]. However, most CAD systems are black-box models. Shen et al. [12] proposed a white-box model with a 3D HSCNN architecture to classify malignant tumors on lung CT images and provide information for the interpretation of decision-making.

Based on ML or DL technology, two recent research articles [13,14] pointed out that many classification models have been proposed since 2016 to diagnose melanoma in skin lesions. For example, Nasr et al. [15] suggested a DL approach with the 8:2 holdout method to achieve 0.810 test accuracy for the MED-NODE data set. Matsunaga et al. [16] used cross-validation to select the best combination of fine-tuning CNN for an ISIC2017 test set after training ResNet-50 with different optimization methods. The AUC efficiency for binary melanoma classification reached 86.8%. Menegola et al. [17] used ResNet-101 and Inception-v4 models to assemble seven well-trained NNs and meta-learning models. The final result obtained the best AUC for the classification of melanoma (87.4%), the third best AUC for the classification of seborrheic keratosis (94.3%), and the third best combined/mean AUC (90.8%) in ISIC 2017 Challenge [18]. Mahbod et al. [19] used an ensemble learning and pre-training network to perform the binary classification of melanoma, thus achieving an AUC efficiency of 87.3%. According to Liu et al. [20], the AUC efficiency of binary classification

of melanoma was 87.0% using a two-stage algorithm of intermediate features. Iqbal et al. [21] proposed a CSLNet architecture and achieved a binary melanoma classification performance with an AUC of 95.2%. By consolidating ISIC 2018 and ISIC 2019 to produce 2299 image sets and utilizing the 8:2 holdout evaluation method, Chang et al. [14] suggested an XGB classifier that incorporated feature extraction and k-means SMOTE techniques to detect melanoma disease, the AUC achieved 98.1%, the F1 score came to 90.5%, and the REC reached 87.8%.

Because the above studies can only provide a black box model for the classification of malignant melanoma disease, it is impossible to understand why the CAD model predicts this dermoscopic image to be malignant. Therefore, González-Díaz [22] proposed a DermakNet architecture, which uses ResNet50 [23] as its core and consists of three subnetworks, to form an entire system. Among them is the lesion segmentation network (Lesion Segmentation Net), which can segment an image into the ROI map of the lesion. A segmentation net of the dermoscopic structure divides each lesion ROI into eight types of dermoscopic structure maps, which are used to interpret diagnostic results. Finally, for disease diagnosis, the diagnostic network (Diagnosis Net) integrates the lesion ROI map and its corresponding dermoscopy structure map. Banerjee et al. [24] used the YOLOv3 algorithm and fuzzy logic to perform the dermoscopic image segmentation task, followed by a pixel-based diagnosis of malignant melanoma. Finally, they performed mathematical operations on an ROI map of the lesion to extract relevant features, such as asymmetry, border, color, and diameter, to explain the diagnostic results. Deep learning and the stacking of machine learning models in groups were used for the diagnosis of skin cancer from melanoma by Alfi et al. [25]. Subsequently, using the SHAP method (shapely adaptive explanations) method, an interpretability approach was constructed that generated heatmaps to identify the regions of an image that were the most suggestive of the disease. Unlike the three studies mentioned above [22,24], in this study, a more direct method was developed that can automatically output diagnostic semantic features and CAD, in addition to AI models for the prediction of malignant melanoma disease.

The research contributions of this study are listed in the following two aspects.

1). This study describes a method for developing an interpretable CNN that dermatologists can use for CAD of malignant melanoma. The intermediate output of the model can predict the diagnostic semantic features related to the final classification of the disease, revealing the decision-making process in the diagnosis of malignant melanoma. To our knowledge, this is the first computer-aided diagnostic AI system with interpretable results in a single architecture for the detection of malignant melanoma.

2). To improve detection performance and obtain diagnostic interpretability, we design various hierarchical network architectures that combine semantic and depth features to predict malignant melanoma. Finally, the hierarchical convolution module of the MEL-HSNet model is chosen and used to learn generalizable features from multiple tasks. The learned information on each characteristic of the skin lesion characteristic is then fed into the final task of predicting malignant melanoma.

The rest of this study is organized as follows. In Section 2, we describe the data set used in this study and the proposed MEL-HSNet model. In Section 3, we present the results and compare the proposed method with other CNN models. In Sections 4 and 5, we discuss the results and conclusions of the study.

2. Materials and methods

2.1. International skin imaging collaboration data set

The International Skin Imaging Collaboration (ISIC) is sponsored by the International Society for Digital Imaging of the Skin (ISDIS) to improve the quality of melanoma diagnosis. The ISIC archive is a public data collection on skin melanoma diseases that they provide; it was used to train and test various methods proposed in this study. The ISIC archive has more than 13,000 dermoscopic images collected from equipment from major international clinical centers. The data source of this study was taken from “ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection” Grand Challenge Data Sets [26,27]. The ISIC 2018 data set is the public data set that was used by ISIC in the 2018 challenge; it includes all the dermoscopic images of multiple types of anatomical parts (excluding mucous membranes and nails). The entire data set comprises three categories: benign, seborrheic keratosis, and melanoma, as well as five image semantic features, including pigment network, negative network, streaks, milia like cyst, and globules. The ISIC 2018 data set provides 2594 training set images and 100 validation set images with ground-truth labels. However, in the validation set, neither the corresponding real labels of the three categories are provided nor the training set provides any corresponding real labels to be used in our evaluation process. Table 1 shows the distribution of positive and negative examples for each attribute in the training, validation, and test sets, where 0 represents the negative example of the category and 1 represents the positive example of the category. Table 2 shows the number distribution for each category in the training, validation, and test sets.

2.2. Our usage of the ISIC2018 data set

Our evaluation required a complete labeled data set, three categories, and five semantic attributes; however, the ISIC 2018 data set does not provide a three-category verification data set, as shown in Tables 1 and 2. Moreover, a complete test data set was required, but the test data set only provided the evaluation results during the challenge period. Therefore, the data used in our entire evaluation process was from the training set of the ISIC 2018 data set, which was divided into 90% (training set) (2335 images) and 10% (test set) (259 images). Furthermore, Tables 1 and 2 show the number of data items for the category of seborrheic keratosis, which is less than the other two categories; therefore, we merged it with the benign category, treated it as a category without melanoma (2075 images), and converted it into a binary melanoma classification task. Although the original three categories were merged into two, the entire data set had an unbalanced data distribution. In particular, the negative network and streaks compared to the other three attributes and the melanoma compared to non-melanoma. After dividing the data into training and test sets, we separately performed data augmentation on the training set [28] to solve the problem of unbalanced data distribution. Table 3 lists the data distribution of each category of the training set, the augmentation set, and the test set after data splitting, merging, and balancing.

Table 1. Detail skin lesion features in the ISIC2018 Task 2 data set.

Feature	Class Type	Training	Validation	Test	Total
Pigment network	0	1071	27	-	1098
	1	1523	73	-	1596
Negative network	0	2404	91	-	2495
	1	190	9	-	199
Streaks	0	2494	94	-	2588
	1	100	6	-	106
Milia like cyst	0	1912	94	-	2006
	1	682	6	-	688
Globules	0	1991	81	-	2072
	1	603	19	-	622
Total		2594	100	-	2694

Table 2. Corresponding skin lesion labels in the ISIC2018 Task 2 data set.

Class Type	Training	Validation	Test	Total
Benign	1867	-	-	1867
Seborrheic Keratosis	208	-	-	208
Melanoma	519	-	-	519
Total	2594	-	-	2594

Table 3. Labels count for the characteristics of the skin lesion.

Feature	Class Type	Training	Augmentation	Test	Total
Pigment network	0	967	1280	104	2351
	1	1368	457	155	1980
Negative network	0	2164	1091	240	3495
	1	171	646	19	836
Streaks	0	2245	960	249	3454
	1	90	777	10	877
Milia like cyst	0	1684	1545	228	3457
	1	651	192	31	874
Globules	0	1775	964	216	2955
	1	560	773	43	1376
Melanoma	0	1868	110	207	2185
	1	467	1627	52	2146
Total		2335	1737	259	4331

2.3. Data preprocessing

The ISIC 2018 data set contains 8-bit RGB dermatoscopy images, and the number of categories is unbalanced. The dermoscopic images of the original data had different image sizes ranging from 771×750 pixels to 6748×4499 pixels. Therefore, to standardize the size of each image

and ensure that the data and memory consumed meet our hardware limitations, we finally resized each dermoscopic image to 224×224 to meet our requirements. As shown in Table 3, the training set created from the ISIC 2018 data set had an imbalance in the number of positives and negatives in each category. To solve this limitation, we performed data augmentation on the training set to balance and increase the amount of training data. In addition, to generate the data size of the 2 to 23 times expanded image set from a small number of original image categories, rotation was performed at various angles for each small number of categories in the training set, such as negative network, streaks, and melanoma, with rotation angles from 10° to 340° . Ultimately, compared to the original un-augmented training data of 2335 images, the augmentation process added a total of 1737 augmented images.

2.4. Hierarchical semantic convolutional neural networks for melanoma diagnosis

In this study, we design a conventional CNN and three interpretable semantic CNNs for the diagnosis of malignant melanoma from dermoscopic images, as shown in Figure 1. They are named MEL-CNN, MEL-HSCNN, MEL-HSMCNN, and MEL-HSNet models. Among them, the MEL-CNN model is a black-box model, whereas the MEL-HSCNN, MEL-HSMCNN, and MEL-HSNet models are all white-box models. First, these models learn the image features through one pre-trained NN model at the beginning of inputting the image. Second, only MEL-HSCNN, MEL-HSMCNN, and MEL-HSNet have a semantic network layer, which contains five variables for the decision-making explanation: pigment network, negative network, streaks, milia-like cyst, and globules. Finally, these models pass the deep features to the final classifier for the classification of melanoma. The detailed architecture of these four models is described below.

In the MEL-CNN model, the popular pre-training model and the global average pooling layer are used as the image feature extractor, and a disease classifier is connected in series, which is designed with a dense layer of 256 nodes and the ReLU activation function, the batch normalization layer, and a dense layer with the sigmoid activation function. The MEL-CNN model is a non-semantic network that uses only image features for melanoma classification.

The architecture of the MEL-HSCNN model consists of an image feature extractor, five semantic classifiers, and a disease classifier. Every semantic classifier contains three dense layers, two batch normalization layers, and a dropout layer. The image feature extraction module is the same as in the MEL-CNN model. The disease classifier incorporates information from both the image features and five semantic features of the dropout layer to predict melanoma.

The MEL-HSMCNN model architecture contains five image feature extractors, five semantic classifiers, and one disease classifier. Compared to the MEL-HSMCNN model, the MEL-HSMCNN model uses the same image feature extractor for the classification of five semantic variables; each semantic classifier in the MEL-HSMCNN model has a separate image feature extraction module.

The architecture of the MEL-HSNet model includes one image feature extractor, five semantic classifiers, and one disease classifier. The image feature extraction module is the same as the MEL-CNN or MEL-HSCNN models. Compared to the MEL-HSMCNN model, the MEL-HSNet model excludes image features and only considers five semantic features of the dropout layer to classify melanoma.

The loss function is one of the most important concepts in machine learning. And optimizing the minimum value of the loss function is the main basis for the model training and learning process. In this study, the above four DL models are trained using binary cross-entropy loss. It should be noted

that the loss function value in each semantic network is accumulated under the same weight to calculate the global loss function value used in this study.

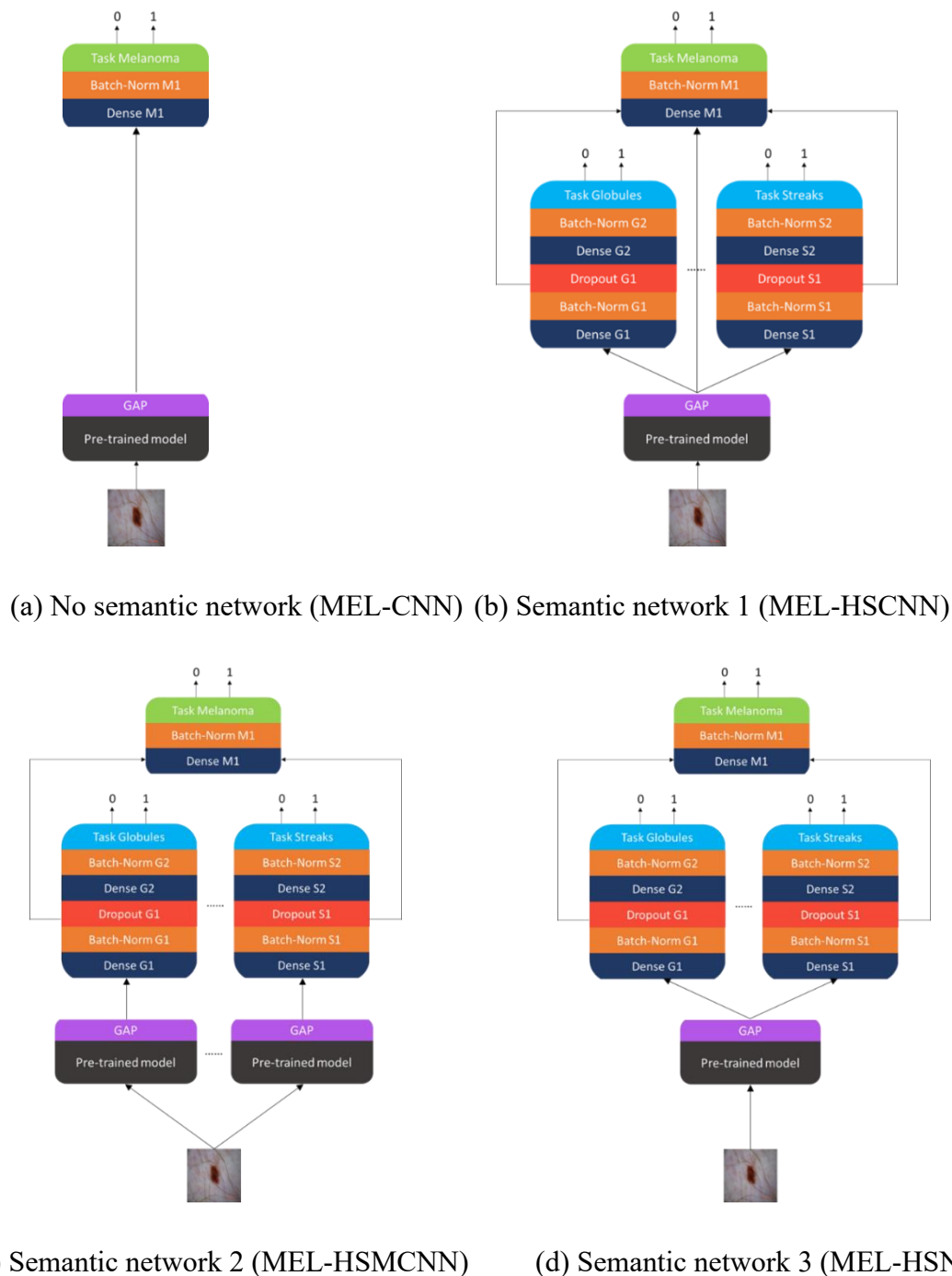


Figure 1. Four new hierarchical semantic convolutional neural networks: (a) No semantic network (MEL-CNN); (b) Semantic network 1 (MEL-HSCNN); (c) Semantic network 2 (MEL-HSMCNN); (d) Semantic network 3 (MEL-HSNet).

2.5. Experimental setup

The four models of MEL-CNN, MEL-HSCNN, MEL-HSMCNN, and MEL-HSNet were

implemented using Python (version 3.8), TensorFlow 2.3 framework, Pandas (version 1.2) library, NumPy (version 1.19) library and Windows 10 version 20H2. The entire training and test process were run on a computer with Nvidia RTX 2060 (6 GB RAM) with CUDA version 10.1 and cuDNN version 7, an AMD Ryzen 5 3600 (3.6 GHz, 6 core CPU), and a main memory of 32 GB. To meet the memory size of the used graphics card, the batch size set during the entire evaluation process was standardized to 16 for the MEL-CNN, MEL-HSCNN, and MEL-HSNet models, and 4 for the MLE-HSMCNN model. The image size was standardized to $224 \times 224 \times 3$ pixels, the data set was divided into 90% for training and 10% for testing, and the experimental results were evaluated over 200 epochs.

2.6. Evaluation measures

In our experimental evaluation process, we used a total of four performance evaluation indicators to evaluate our experimental results: sensitivity (SEN), specificity (SPE), accuracy (ACC), and area under the curve (AUC). True positive (TP) was used when calculating certain formulas, where the actual positive samples are predicted as positive; true negative (TN): the actual negative samples are predicted as negative; false positive (FP): the actual negative samples are predicted as positive; and false negative (FN): the actual positive samples are predicted as negative.

Sensitivity refers to the proportion of samples that are positive and predicted to be positive as follows:

$$SEC = \frac{TP}{TP+FN} \quad (1)$$

Specificity refers to the proportion predicted to be negative in an actual negative sample as follows:

$$SPE = \frac{TN}{TN+FP} \quad (2)$$

Accuracy refers to the prediction of the correct scale in all the actual samples as follows:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The AUC of the ROC denotes a positive and negative sample selected at random, and the classifier correctly set the positive sample with a higher probability of scoring than the negative sample as follows:

$$AUC = \int_0^1 T(F_0) dF_0 \quad (4)$$

where $T(F_0)$ is the corresponding true positive rate when the false positive rate is F_0 .

3. Results

In this section, we explain the several models we used, the fine-tuning of the different architectures of the models, and the results of comparing them under different conditions. Subsequently, we evaluate the prediction accuracy of our diagnostic semantic features and provide explanations for the correct and incorrect predictions.

3.1. Effect of the imbalanced vs. balanced training data

Imbalanced data means an unequal distribution of classes within a data set. This is a common

occurrence in medical data sets. To balance the data set, we used a data augmentation technique to augment the original data and increase the training data from the original 2335 to 4072. This experiment used ResNet50 as the pre-training model and embedded MEL-HSNet as the classification model to classify the data to understand whether balanced data have different effects on the image classification results. Table 4 lists the results of the use of unbalanced and balanced data. The bold values in the table represent the results of the better model. Compared to using unbalanced data as training data, the sensitivity of balanced data for melanoma classification increased by 17.3% and the AUC value increased by 6.43%. Balancing of the data could effectively improve the performance of the model; therefore, balanced data was used as training data in the subsequent experiments.

Table 4. Classification performance (%) of the test set for different training data distributions.

Training Data	SEN	SPE	ACC	AUC
Imbalanced	63.46	95.65	89.18	87.02
Balanced	80.76	97.10	93.82	93.45

3.2. Effect of with vs. without the pre-trained model

Standing on the shoulders of giants is a common metaphor for using a pre-trained model. Using the learned feature maps of these pre-trained models, we can quickly obtain better results without the need to start from scratch by training a model. Therefore, in the subsequent experiments, after deciding to use balanced data as training data, we aimed to understand whether using a pre-trained model would influence the classification performance of our model. This experiment used balanced data as training data and MEL-HSNet as a classification model to test different conditions. For the experiment, we used the simple feature model constructed by us, InceptionResNetV2 and ResNet50, as different feature models for testing. Moreover, we fine-tuned the pre-trained model, and the results are shown in Table 5. The bold values in the table represent the results of the better model. The first two results demonstrate that when using the pre-trained model InceptionResNetV2 as the feature model, the sensitivity and AUC values significantly improved from 13.46 and 81.85% to 71.15 and 90.94%, respectively, compared with the feature model we built. Moreover, after using ResNet50 as the feature model as presented in the last row, the sensitivity and AUC values reached 80.76 and 93.45%, respectively. As shown in Table 5, the ResNet50 pre-trained model outperformed the non-pre-trained model and the InceptionResNetV2 pre-trained model; therefore, as the feature learning model, we decided to use the ResNet50 pre-trained model in the subsequent experiments.

Table 5. Classification performance (%) of the test set for different pre-trained models.

Pre-trained Model	SEN	SPE	ACC	AUC
MEL-HSNet (w/o Pre-trained Model)	13.46	99.03	81.85	70.88
MEL-HSNet (w/ InceptionResNetV2)	71.15	95.65	90.73	90.94
MEL-HSNet (w/ ResNet50)	80.76	97.10	93.82	93.45

3.3. Effect of different semantic models

In this experiment, our objective was to know whether our modified MEL-HSNet model

architecture allows the semantic network to provide different layers of results (that is, the concatenate layer) to the final melanoma classification layer and whether a better model architecture than the original results can be achieved. We tested three architectures that use each semantic network as the final concatenate and provided input information to the melanoma classification layer. They are as follows: 1) activation layer, 2) batch normalization layer, and 3) dropout layer. Furthermore, the classification performance of MEL-HSNet, MEL-CNN, MEL-HSCNN, and MEL-HSMCNN models was compared. Table 6 lists the results of the final comparison. The MEL-HSNet model with the dropout layer version gave the best results among the three architectures. The sensitivity, specificity, accuracy and AUC values were 80.76, 97.10, 93.82, and 93.45%, respectively. Compared to MEL-CNN, the sensitivity was improved by 11.53%. Compared with MEL-HSCNN and MEL-HSMCNN, the MEL-HSNet model performed better, in addition to reducing training costs in terms of time and memory. Therefore, we focused on using the MEL-HSNet (concat Dropout) model in the subsequent experiments.

Table 6 shows that the MEL-HSMCNN model had the highest capacity, indicating that the model had the longest training time. The bold values in this table represent the results of a better model. The MEL-CNN model had the smallest capacity; therefore, the model had the shortest training time. In this study, the proposed MEL-HSNet model (concat Dropout) has a capacity and training time similar to the MEL-CNN model.

Table 6. Classification performance (%) of the test set for different semantic models.

Classification Model w/ ResNet50	SEN	SPE	ACC	AUC	CPU (s)	MEM (mb)
MEL-CNN	69.23	99.51	93.43	94.87	6802	24.11
MEL-HSCNN	71.15	95.65	90.73	91.75	7119	24.87
MEL-HSMCNN	63.46	97.10	90.34	91.55	54,430	18.70
MEL-HSNet (concat Activation)	69.23	98.06	92.27	95.81	7091	24.94
MEL-HSNet (concat Batch-Norm)	76.92	95.65	91.89	94.07	7156	24.94
MEL-HSNet (concat Dropout)	80.76	97.10	93.82	93.45	7109	24.94

3.4. Melanoma prediction performance

In this section, we demonstrate the overall classification performance of our model. Balanced data was used as training data, and ResNet50 was used as a pre-trained model in feature learning. The dropout layer version of MEL-HSNet was our final model, and we compared it with MEL-CNN. Table 7 lists the results of the comparison. The bold values in the table represent the results of the better model. The overall results of MEL-HSNet and MEL-CNN were improved after using ResNet50. Finally, MEL-HSNet achieved a sensitivity of 80.76%, a specificity of 97.10%, an accuracy of 93.82% and an AUC value of 93.45%. MEL-CNN had a sensitivity of 69.23%, a specificity of 99.51%, an accuracy of 93.43%, and an AUC value of 94.87%. Figure 2 shows the confusion matrix for melanoma classification using MEL-CNN and MEL-HSNet with or without the pre-trained model. The results show that using the ResNet50 pre-trained model, MEL-HSNet could better classify the positive and negative melanoma cases compared with other models. The receiver operating characteristic (ROC) curve was compared with the four classification models, as shown in Figure 3. The result in Figure 3 demonstrates that MEL-HSNet with ResNet50 pre-trained

model outperforms the other three models.

The confusion matrix, the ROC map and the metric evaluation show that the MEL-HSNet method using semantic networks for classification performed better in predicting melanoma compared to the MEL-CNN method using features from the learned images directly.

Table 7. Classification performance (%) of the test set for melanoma prediction.

Classification Model	SEN	SPE	ACC	AUC
MEL-CNN (w/o Pre-trained Model)	26.92	97.58	83.39	80.49
MEL-CNN (w/ ResNet50)	69.23	99.51	93.43	94.87
MEL-HSNet (w/o Pre-trained Model)	13.46	99.03	81.85	70.88
MEL-HSNet (w/ ResNet50)	80.76	97.10	93.82	93.45

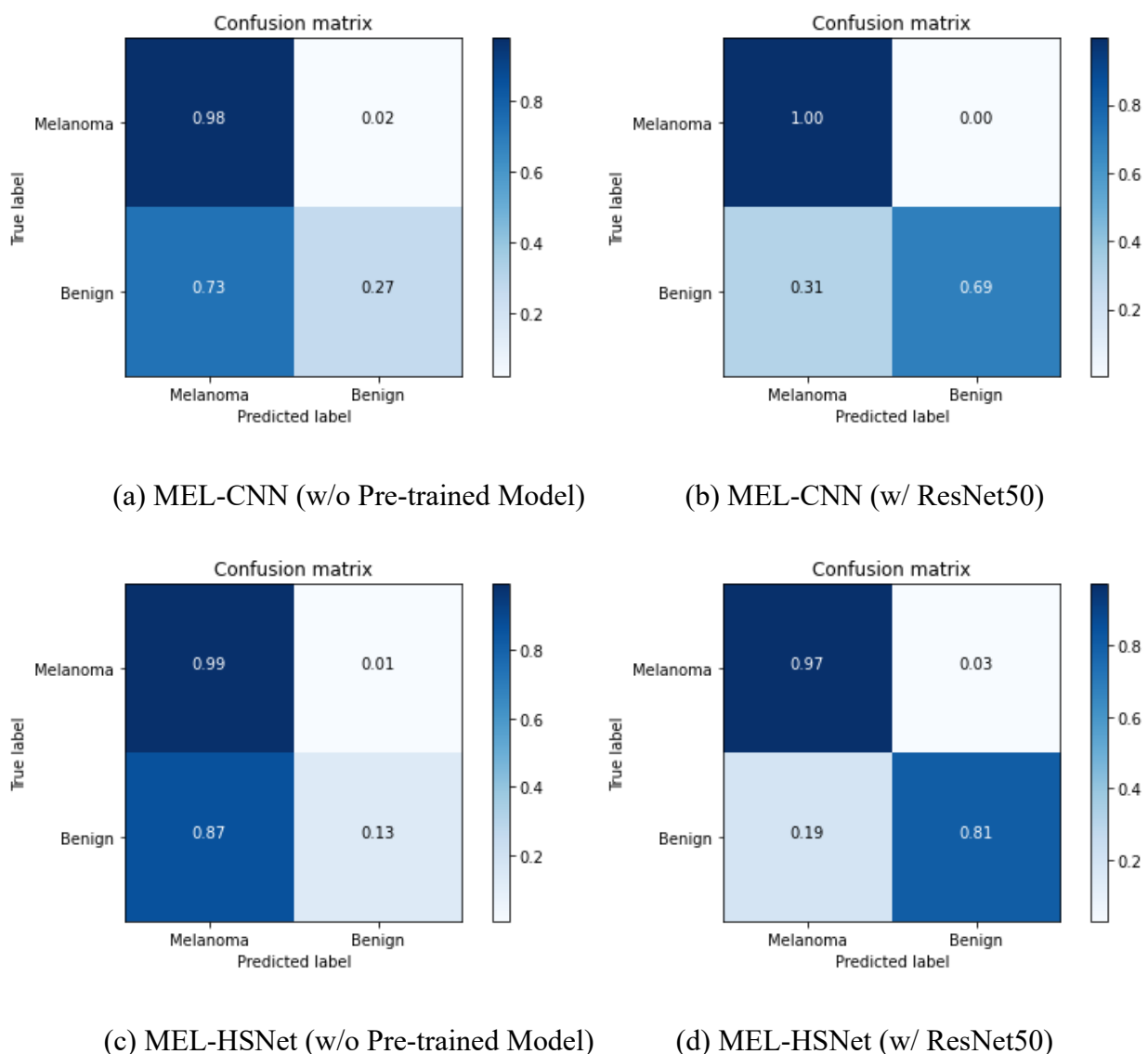


Figure 2. Confusion matrices of the test set for melanoma prediction: (a) MEL-CNN (w/o Pre-trained Model); (b) MEL-CNN (w/ ResNet50); (c) MEL-HSNet (w/o Pre-trained Model); (d) MEL-HSNet (w/ ResNet50).

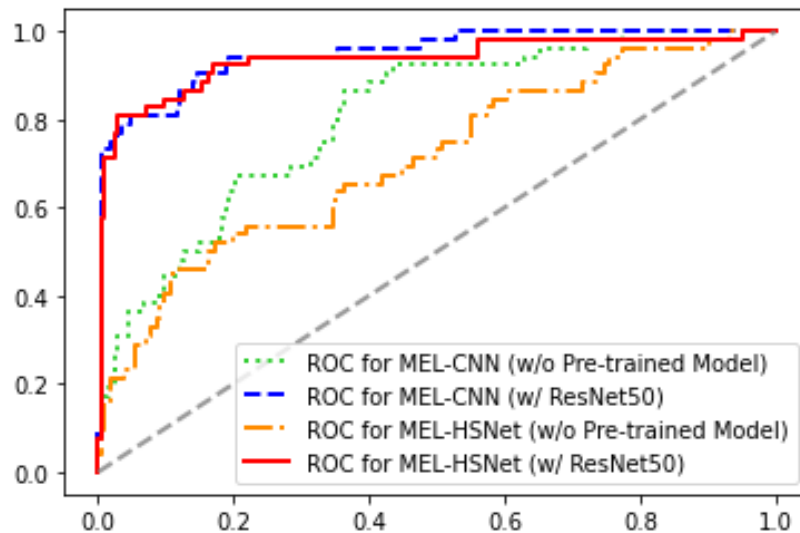


Figure 3. Receiver operating characteristic curve comparison.

In addition to demonstrating the classification performance of the model in melanoma, we presented the results of the five semantic features predicted by the MEL-HSNet semantic network using ResNet50. For each diagnostic semantic feature in Table 8, the accuracy rates were 74.13, 92.66, 96.91, 83.01, and 82.62% for the pigment network, negative network, streaks, milia-like cyst, and globules, respectively. Moreover, the AUC values were 82.97, 71.38, 86.70, 62.79, and 73.24%. The results demonstrate that the MEL-HSNet model can achieve good results in predicting the five semantic features it contains, as well as in classifying melanoma.

Table 8. Classification performance (%) of the test set for the semantic feature predictions.

Semantic Features	ACC	AUC
Pigment network	74.13	82.97
Negative network	92.66	71.38
Streaks	96.91	86.70
Milia like cyst	83.01	62.79
Globules	82.62	73.24

3.5. Semantic feature prediction and model interpretability

Finally, in this section, we discuss semantic network classification. Figure 4 demonstrates the interpretability of the MEL-HSNet model by presenting dermoscopic images, interpretable semantic labels for image prediction, and classification results for melanoma. As shown in Figure 4, the MEL-HSNet model predicts three dermoscopic images as benign or malignant, which is the same as the actual labels. Furthermore, the predicted semantic features are also quite close to the actual labels. Therefore, we conclude that MEL-HSNet can predict the results with six semantic features simultaneously, providing better interpretability compared to the MEL-CNN model.


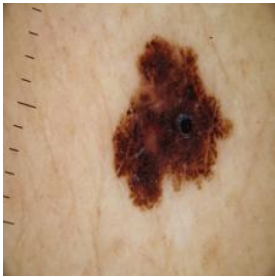
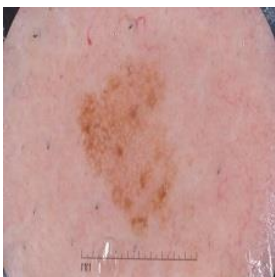
	Semantic Features	True Label	Prediction
	Pigment network	1	0
	Negative network	0	0
	Streaks	0	0
	Milia like cyst	1	1
	Globules	0	0
	Melanoma	0	0
	Semantic Features	True Label	Prediction
	Pigment network	1	1
	Negative network	0	0
	Streaks	0	0
	Milia like cyst	0	0
	Globules	0	0
	Melanoma	1	1
	Semantic Features	True Label	Prediction
	Pigment network	0	0
	Negative network	0	0
	Streaks	0	0
	Milia like cyst	0	1
	Globules	0	0
	Melanoma	1	1

Figure 4. Illustration of the MEL-HSNet (w/ ResNet50) model's interpretability.

4. Discussion

Recently, studies demonstrated that DL methods perform better in terms of the binary classification of skin melanoma [16,17,19–21]. However, since these models are black-box models, they can only map image features into a category prediction for skin lesions, and therefore cannot provide sufficient explanatory power for medical diagnosis. In this study, the three developed models (MEL-HSCNN, MEL-HSMCNN, and MEL-HSNet) are white-box models that could simultaneously predict the malignancy of skin melanoma and detect five characteristics of the skin lesion.

I. Gonzalez-Diaz [22] proposed a CAD system called DermakNet, which uses the Dermoscope Structure Segmentation Network (DSSN) subsystem to achieve the interpretability of its diagnosis. DSSN used a constrained CNN for lesion segmentation as a pixel-wise labeling problem, and provided a pixel-based dermoscopic feature map, which can be understood and interpreted by the human eye. In this study, the proposed MEL-HSNet architecture used the hierarchical semantic CNN to directly predict five diagnostic semantic feature labels for malignant melanoma. Moreover, the interpretation mechanism of the proposed MEL-HSNet model was relatively concise and convenient.

S. Banerjee et al. [24] used the YOLOv3 algorithm and fuzzy logic to diagnose malignant

melanoma in a pixel-based manner. The ROI map of the skin lesion was then subjected to mathematical operations to extract four features that can be used to diagnose melanoma with traditional ABCD (Asymmetry, Border irregularity, Color variation, and Diameter) clinical guidelines. Compared to the interpretation mechanism of S. Banerjee et al. [24], the MEL-HSNet architecture proposed in this study is a CAD model that can automatically and synchronously produce five diagnostic semantic features and malignant melanoma prediction scores.

I. A. Alfi et al. [25] used SHAP to create heatmaps to identify which regions of an image are more associated with melanoma disease. As with SHAP, the heatmaps of the four representative images generated by GradCAM++ [29] are shown in Figure 5 in this study. In contrast to the MEL-HSNet model proposed here, using post hoc explanatory methods such as SHAP or GradCAM can only tell the user which regions in the image are most relevant to the class predicted by the model. The explanation mechanism in this study is to inform users of disease prediction results and the corresponding five semantic features of the image, as shown in Figure 4.

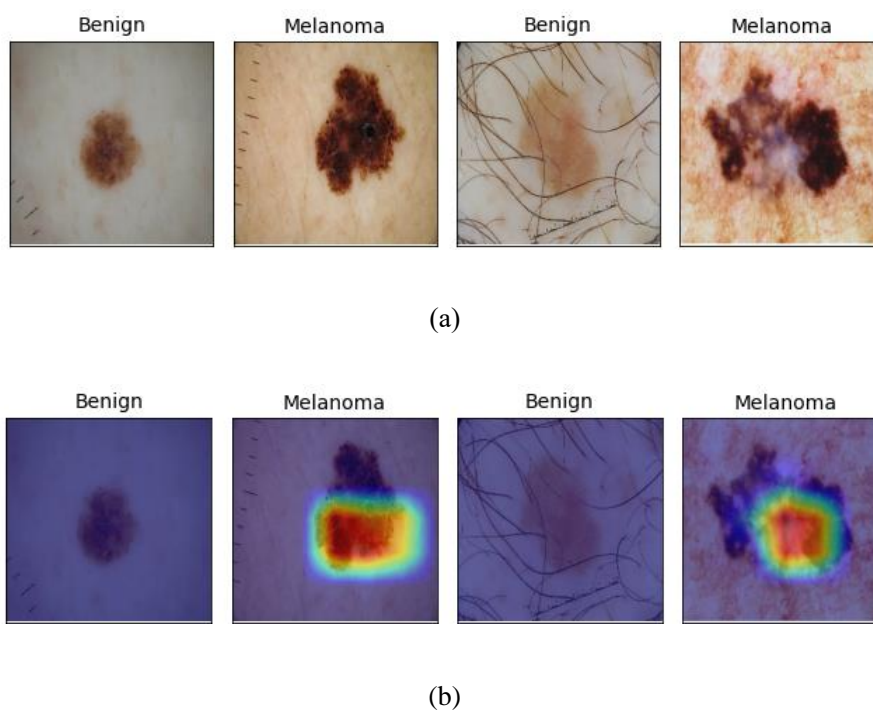


Figure 5. Representative cases on ISIC2018 data set: (a) original image; (b) explanation by GradCAM++ analysis.

We also compare the performance of our proposed model with the SOTA approaches for the binary classification of melanoma published recently. Based on the dataset, the classification method, the validation method, and the performance of the test set, a comparative summary of these methods is provided in Table 9. Since different studies use different data sizes, valid comparisons are difficult. However, the method proposed in this study still exhibits excellent performance.

Table 9. A comparative summary of the SOTA approaches for the binary classification of melanoma.

Year	Author	Dataset	Method	Validation	Test Result
2020	J. A. Almaraz-Damian et al. [30]	ISIC 2018	DL+ML	Holdout (75:25) full: 10015	ACC: 0.897
2020	J. Daghrir et al. [31]	Subset of ISIC archive	DL+ML	Holdout (8:2) full: 640	ACC: 0.884
2022	I. A. Alfi et al. [25]	Subset of ISIC 2018	ML	Holdout (8:2) full: 3297	ACC: 0.880
2022	I. A. Alfi et al. [25]	Subset of ISIC 2018	DL	Holdout (8:2) full: 3297	ACC: 0.910 AUC: 0.970
2023	Our approach MEL-HSNet	ISIC 2018 Task 2	DL	Holdout (9:1) full: 4331	ACC: 0.938 AUC: 0.935

There are three limitations to this study. The most significant limitation of this study is the small data set. The number of dermoscopic images with skin lesion features and malignant melanoma labels is limited; only 2594 ground truth cases (see Table 3) were obtained in this study. Additionally, the entire data set is divided into the training and test sets. The training set is used to train the model, and the test set is used to evaluate the model. Due to the lack of a validation set for the verification of model hyperparameters, this study relied on the relevant setting values from the previous literature [16,17,19–22,24], and the early stopping criterion [32] could not be used to improve overall model generalization. Finally, the disease category and the corresponding five diagnostic semantic features of this study were scaled to binary labels. These limitations can be improved by modeling large annotated data sets that contain discriminatory features. Therefore, the topic of model optimization remains to be investigated in the future.

5. Conclusions

Malignant melanoma, known as melanoma, is a type of skin cancer that is more dangerous than other types of skin cancer because it metastasizes quickly if not diagnosed and treated in its early stage. Recently, based on DL models, some studies have succeeded in achieving extremely promising results. However, most of their models are black-box models. It is crucial to provide an interpretation ability of a computer-aided diagnosis model for the proper detection of such fatal skin diseases.

This study focused on the design and evaluation of interpretable deep network frameworks for melanoma CAD such as MEL-CNN, MEL-HSCNN, MEL-HSMCNN, and MEL-HSNet. After

performance analysis, we proposed the MEL-HSNet model, which can predict the melanoma score while classifying five semantic features of the melanoma (pigment network, negative network, streaks, milia-like cyst, and globules).

Although the results of the performance evaluation of the MEL-HSNet model proposed in this study are encouraging and may provide a promising future application to help dermatologists diagnose malignant melanoma, more clinical cases need to be collected to optimize the model to meet the requirements of future clinical practice.

Acknowledgments

This study was partially funded by the National Science and Technology Council of Taiwan, grant number MOST 111-2121-M-040-001.

Conflict of interest

All authors declare no any potential financial and non-financial conflicts of interest.

References

1. F. Bray, M. Laversanne, E. Weiderpass, I. Soerjomataram, The ever-increasing importance of cancer as a leading cause of premature death worldwide, *J. Cancer*, **127** (2021), 3029–3030. <https://doi.org/10.1002/cncr.33587>
2. WHO, Global health estimates 2020: deaths by cause, age, sex, by country and by region, 2000–2019, 2020. Available from: <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death>.
3. S. Carr, C. Smith, J. Wernberg, Epidemiology and risk factors of melanoma, *Surg. Clin. North Am.*, **100** (2020), 1–12. <https://doi.org/10.1016/j.suc.2019.09.005>
4. D. S. Rigel, J. A. Carucci, Malignant melanoma: prevention, early detection and treatment in the 21st century, *CA: Cancer J. Clin.*, **50** (2000), 215–236. <https://doi.org/10.3322/canjclin.50.4.215>
5. I. H. Wolf, J. Smolle, H. P. Soyer, H. Kerl, Sensitivity in the clinical diagnosis of malignant melanoma, *J. Melanoma Res.*, **8** (1998), 425–429. <https://doi.org/10.1097/00008390-199810000-00007>
6. M. L. Bafounta, A. Beauchet, P. Aegerter, P. Saiag, Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests, *Arch. Dermatol.*, **137** (2001), 1343–1350. <https://doi.org/10.1001/archderm.137.10.1343>
7. H. P. Soyer, G. Argenziano, R. Talamini, S. Chimenti, Is dermoscopy useful for the diagnosis of melanoma, *Arch. Dermatol.*, **137** (2001), 1361–1363. <https://doi.org/10.1001/archderm.137.10.1361>
8. M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, et al., A methodological approach to the classification of dermoscopy images, *Comput. Med. Imaging Graphics*, **31** (2007), 362–373. <https://doi.org/10.1016/j.compmedimag.2007.01.003>

9. C. Barata, M. E. Celebi, J. S. Marques, A survey of feature extraction in dermoscopy image analysis of skin cancer, *IEEE J. Biomed. Health. Inf.*, **23** (2019), 1096–1109. <https://doi.org/10.1109/JBHI.2018.2845939>
10. T. Y. Wong, N. M. Bressler, Artificial intelligence with deep learning technology looks into diabetic retinopathy screening, *JAMA*, **316** (2016), 2366–2367. <https://doi.org/10.1001/jama.2016.17563>
11. A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, H. J. W. L. Aerts, Artificial intelligence in radiology, *Nat. Rev. Cancer*, **18** (2018), 500–510. <https://doi.org/10.1038/s41568-018-0016-5>
12. S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, W. Hsu, An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification, *Expert Syst. Appl.*, **128** (2019), 84–95. <https://doi.org/10.1016/j.eswa.2019.01.048>
13. D. Popescu, M. El-Khatib, H. El-Khatib, L. Ichim, New trends in melanoma detection using neural networks: a systematic review, *Sensors*, **22** (2022), 496. <https://doi.org/10.3390/s22020496>
14. C. C. Chang, Y. Z. Li, H. C. Wu, M. H. Tseng, Melanoma detection using XGB classifier combined with feature extraction and K-means SMOTE techniques, *Diagnostics*, **12** (2022), 1747. <https://doi.org/10.3390/diagnostics12071747>
15. E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, et al., Melanoma detection by analysis of clinical images using convolutional neural network, in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, **2016** (2016), 1373–1376. <https://doi.org/10.1109/EMBC.2016.7590963>
16. K. Matsunaga, A. Hamada, A. Minagawa, H. Koga, Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble, preprint, arXiv:1703.03108.
17. A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, E. Valle, RECOD titans at ISIC challenge 2017, preprint, arXiv:1703.04819.
18. A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, E. Valle, RECOD Titans participation at the ISBI 2017 challenge - Part 3. Available from: <https://github.com/learningtitans/isbi2017-part3>.
19. A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, C. Wang, Fusing fine-tuned deep features for skin lesion classification, *J. Comput. Med. Imaging Graphics*, **71** (2019), 19–29. <https://doi.org/10.1016/j.compmedimag.2018.10.007>
20. L. Liu, L. Mou, X. X. Zhu, M. Mandal, Automatic skin lesion classification based on mid-level feature learning, *Comput. Med. Imaging Graphics*, **84** (2020), 101765. <https://doi.org/10.1016/j.compmedimag.2020.101765>
21. I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, J. Ma, Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images, *Comput. Med. Imaging Graphics*, **88** (2021), 101843. <https://doi.org/10.1016/j.compmedimag.2020.101843>
22. I. Gonzalez-Diaz, Dermaknet: incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis, *IEEE J. Biomed. Health Inf.*, **23** (2018), 547–559. <https://doi.org/10.1109/JBHI.2018.2806962>
23. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. <https://doi.org/10.1109/CVPR.2016.90>

24. S. Banerjee, S. K. Singh, A. Chakraborty, A. Das, R. Bag, Melanoma diagnosis using deep learning and fuzzy logic, *Diagnostics*, **10** (2020), 577. <https://doi.org/10.3390/diagnostics10080577>
25. I. A. Alfi, Md. M. Rahman, M. Shorfuzzaman, A. Nazir, A non-invasive interpretable diagnosis of melanoma skin cancer using deep learning and ensemble stacking of machine learning models, *Diagnostics*, **12** (2022). <https://doi.org/10.3390/diagnostics12030726>
26. N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, et al., Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC), preprint, arXiv:1902.03368.
27. P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Sci. Data*, **5** (2018), 180161. <https://doi.org/10.1038/sdata.2018.161>
28. C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data*, **6** (2019), 60. <https://doi.org/10.1186/s40537-019-0197-0>
29. A. Chattopadhyay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks, in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2018. <https://doi.org/10.1109/WACV.2018.00097>
30. J. A. Almaraz-Damian, V. Ponomaryov, S. Sadovnychiy, H. Castillejos-Fernandez, Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures, *Entropy*, **22** (2020). <https://doi.org/10.3390/e22040484>
31. J. Daghbir, L. Tlig, M. Bouchouicha, M. Sayadi, Melanoma skin cancer detection using deep learning and classical machine learning techniques: a hybrid approach, in *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, IEEE, (2020), 1–5. <https://doi.org/10.1109/ATSIP49331.2020.9231544>
32. R. Caruana, S. Lawrence, L. Giles, Overfitting in neural nets: backpropagation, conjugate gradient and early stopping, in *Advances in Neural Information Processing Systems*, (2001), 402–408. Available from: <https://www.researchgate.net/publication/221620260>.



AIMS Press

©2023 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)