



A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation

Onur Sevli¹

Received: 28 November 2020 / Accepted: 12 March 2021 / Published online: 24 March 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

Pigmented skin lesions are common spots or growths on the skin that originate from melanocyte cells. Skin cancer occurs as a result of the uncontrolled division of melanocyte cells. Skin cancer is common worldwide and its incidence has been increasing. Timely and accurate diagnosis of skin cancer is important in reducing deaths. Skin cancer cases are diagnosed by expert dermatologists, but the number of experts is insufficient for the population. Dermoscopy is one of the most commonly used methods for imaging skin lesions. Sometimes misdiagnosis can occur in the interpretation of dermoscopic images by human experts. Computer-aided systems help to make accurate and objective decisions in the diagnosis of skin lesions. Convolutional neural network(CNN) is a deep learning technique commonly used in the field of computer vision and is also widely used in medical image analysis. In this study, a CNN model was proposed to classify seven different skin lesions in the HAM10000 dataset. The model provided a classification accuracy of 91.51%. The performance of the model was compared with similar studies in the literature and it was found that it showed higher success than most studies. The model was connected to a web application and evaluated in two phases by seven expert dermatologists. In the first phase, it was concluded that the model could diagnose skin lesions with 90.28% accuracy in practice. In the second phase, the model corrected the experts' misdiagnoses by 11.14%.

Keywords Skin lesion classification · Melanoma detection · Convolutional neural network · HAM10000 dataset

1 Introduction

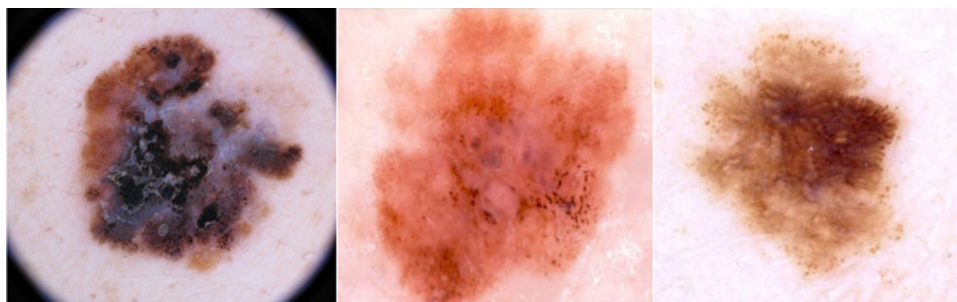
Pigmented skin lesions are common spots and growths of different colors and shapes on the skin. These formations originate from cells called melanocytes which produce melanin pigment that gives color to the skin. Birthmarks, age spots, freckles, and moles are all pigmented skin lesions. Although most of them are harmless, some pigmented lesions carry the risk of skin cancer. Cancer is caused by abnormal and uncontrolled cell division. The uncontrolled division of melanocyte cells causes skin cancer known as melanoma. Malignant melanomas with the risk of spreading are asymmetrical, irregular, and rough lesions with a diameter greater than 5 mm and, with

multiple color tones (Fig. 1). Genetic predisposition, prolonged exposure to sun or artificial ultraviolet light sources [1], ionizing and infrared radiation, human papillomavirus (HPV), and smoking [2] are also among the risk factors for malignant melanomas.

Skin cancer is a common type of cancer worldwide and its incidence has increased in the last decade [3]. Skin cancer is considered in two types as melanoma and non-melanoma. Nonmelanoma skin cancer occurs in the upper layers of the skin and grows slowly. Melanoma skin cancer occurs in any part of the body, especially on the back in men, and on the legs in women, and tends to spread to other organs and tissues [4]. Melanoma skin cancer is less common than nonmelanoma but is more fatal. According to the statistics of the World Health Organization (WHO), around 3 million new nonmelanoma and 132 thousand new melanoma cases have been seen worldwide every year. The number of cases is increasing with each passing year and one of the main reasons for this increase is more exposure to the harmful effects of the sun [5]. According to the

✉ Onur Sevli
onursevli@mehmetakif.edu.tr

¹ Computer Engineering Department, Faculty of Engineering and Architecture, Burdur Mehmet Akif Ersoy University, 15030 Burdur, Turkey

Fig. 1 Malignant skin lesions

statistics of the American Cancer Society, a 53% increase in melanoma cases was observed between 2008 and 2018. Early diagnosis and treatment of melanomas are crucial to reducing deaths. There is a need for trained experts in the diagnosis of skin cancer, but the number of experts is not sufficient compared to the number of cases. In 2018, there were only 9600 dermatologists for 323 million people in the USA [6]. The fact that the number of experts is insufficient makes it necessary to develop new solutions for fast and effective diagnosis.

Medical imaging provides great convenience in cancer diagnosis. Dermatoscope is equipment that displays skin lesions in high resolution and is used in the diagnosis of skin cancer. Dermoscopy, an examination using a dermatoscope, increases the success of diagnosing benign and malignant skin lesions. Studies show that diagnoses made on dermoscopic images give more accurate results [7].

Despite special training and dermoscopic analysis, dermatologists tend to misdiagnose up to 20% in clinical tests. These misdiagnoses are caused by human factors such as fatigue and mental state [8]. Computer-aided systems help to make objective, fast and reliable decisions in the diagnosis of skin lesions as in many areas [9]. Today, advances in artificial intelligence technologies and successful studies in the field of computer vision have led to a transformation in the field of medicine [10].

Computer-aided dermoscopic image analysis studies consist of determining the area where the lesion is located, extracting the lesion features, and classification stages [11]. Besides, pre-processings such as contrast adjustment [12], color balancing [13], normalization [14], and calibration of the image [15] are also applied. Computer-aided dermoscopic analysis systems help experts make fast and accurate decisions [16].

Fatima et al. conducted a study that detected melanomas by statistically analyzing 21 pre-determined parameters of skin lesions using a computer-assisted six-step method [17]. Qian et al. with a two-stage method, first determined the location of the lesion and then classified it. They stated that this two-stage method was more successful than a single-stage classification [18]. Erçal et al. conducted a study that distinguishes malignant melanomas from 3

different benign lesions with similar structures, with 80% success using artificial neural networks [19]. In another study, Erçal et al. proposed the use of fuzzy logic and hierarchical neural networks for skin lesion classification [20]. Binder et al. compared the success of computer-aided systems and human experts in classifying benign and malignant tumors to test the effectiveness of artificial neural networks in dermoscopic image analysis. In their study, which was carried out with a hundred images for both classes, human experts achieved 88% accuracy and the artificial neural network achieved 86% accuracy [21]. Pereira et al. achieved 77% and 78% accuracy in their classification study with support vector machine (SVM) and feed-forward network on a dataset consisting of 170 benign and malignant skin lesion images [22]. Almaraz-Damian et al. using handcrafted features achieved 92% success in classifying benign and malignant skin lesions with linear regression and SVM [23]. In another study, Kavitha and Suruliandi extracted texture and color features from a dataset of 150 dermoscopic images and then classified melanoma and non-melanoma lesions with 93% accuracy using SVM [24]. However, the studies mentioned above were carried out with small datasets and therefore their generalizability was low.

Deep learning, a sub-branch of machine learning, provides automatic extraction of complex feature hierarchies within large datasets with the help of multi-layered neural networks. Deep learning can perform many challenging tasks with high success, such as analysis of complex patterns without the need for manual feature extraction. For this reason, deep learning has become more popular in medical image analysis studies. Studies were shown that deep learning methods provide successful results in applications for cancer diagnosis [25]. In the field of skin cancer diagnosis, studies have been carried out to detect malignant tumors with deep learning [26]. In recent years, Convolutional Neural Network (CNN), a deep learning technique, is widely used in medical image processing [27].

Mahbod et al. proposed a CNN model that classifies melanoma with 87% accuracy over 600 lesion images [28]. In another study, Kawahara et al. achieved 81% success in classifying 1300 skin lesion images taken with a non-

dermoscopic camera [29]. Zhang et al. achieved 85% accuracy in melanoma classification using the ARL-CNN50 model [30]. Bi et al. achieved 74% accuracy with a hyper-connected CNN model to classify a dataset consists of 1011 skin lesion images [31]. In a study conducted by Ayan and Üner with a limited dataset, it was concluded that data augmentation increases the model performance [32]. Menegola et al. conducted a study demonstrating that transfer learning increases success in classifying skin lesions [33]. Lopez et al. reached 78% accuracy in their study for early detection of melanomas via transfer learning [34]. Mahbood et al. reached an accuracy of 86% on the HAM10000 dataset with the transfer learning method [35]. Gessert et al. achieved an AUC for sensitivity above 80% in the lesion classification they performed with a pre-trained CNN model [36]. Yilmaz and Trocan achieved an accuracy of 91%, 92%, and 92% in their benign and malignant skin lesion classification study, which they performed with 3 different pre-trained CNN models [37]. Li and Shen achieved 91.2% accuracy using FCRN deep convolutional neural network and data augmentation, on a dataset containing 2000 skin lesion images of three different types [38]. Salian et al., reached an accuracy of 80% in their study with Mobilenet, VGG-16, and a custom CNN model, on a dataset consisting of 200 skin lesion images of 6 different classes [39]. Serte and Demirel achieved an accuracy of 81% and 83% in the classification of 2000 skin lesion images with pre-trained AlexNet and ResNet-18 [40]. Liu et al. achieved an average accuracy of 87% in the benign and malignant skin lesion classification with DenseNet and ResNet models [41]. Milton achieved the highest accuracy of 76% in a comparative skin lesion classification study performed with pre-trained InceptionResNetV2, SENet154, InceptionV4 models [42]. Aldwgeri and Abu-backer obtained accuracy values varying between 74 and 77% in their classification study using pre-trained VGG-Net, ResNet50, InceptionV3, Xception, and DenseNet121 models [43]. Anjum et al. achieved 93% accuracy in their study with pre-trained models and YOLO on the HAM10000 skin lesion images dataset [44]. In different studies carried out with the transfer learning method, the researchers achieved an accuracy of up to 82% by using various models such as Vgg, AlexNet, DenseNet, ResNet, Inception, Xception on large datasets [45–48].

In this study, a CNN model is proposed for pigmented skin lesion classification. Images of skin lesions consisting of 10,015 samples and seven classes were classified with this model. The performance of the model was characterized by accuracy, precision, recall, and F1 score. The trained model was connected to a web application via a REST-API. The application was tested by expert dermatologists. As a result of the tests carried out in two phases, it was concluded that the developed model is a successful

decision support system according to the feedback of the experts and can be applied in practice after further tests.

2 Materials and methods

2.1 The dataset

Because the image sets for the diagnosis of pigmented skin lesions are small and inadequate, and deep learning methods require larger datasets to produce healthy results, Tschandl et al. created the HAM10000 (Human Against Machine with 10,000 training images) dataset with dermoscopic images obtained from different populations [49]. This dataset contains all the important diagnostic categories of pigmented lesion types. The current version of the dataset can be accessed through the International Skin Imaging Collaboration (ISIC) archive [50]. The dataset served as the training set for the ISIC 2018 challenge was shared publicly as a resource for self-diagnosis and decision support systems. The main purpose of the dataset is to support applications that provide an accurate diagnosis of skin lesions, thereby reducing deaths from melanoma.

It is possible to group the lesions encountered in clinical practice into seven different types [49]. In this dataset, there are 10,015 dermoscopic images of these seven types. The diagnoses of the skin lesions in the dataset were made by expert dermatologists. The seven categories in the dataset are:

- Actinic keratosis (akiec): It is a type of epidermal lesion that occurs in sensitive areas of the body exposed to sunlight for a long time. Commonly, their surface looks like a rough patch. The risk of this type of lesion turning into cancer is low. These lesions, starting with 1 to 2 mm in size and reaching larger sizes, can be treated locally without the need for surgical operation [51].
- Basal cell carcinoma (bcc): It is a type of skin cancer that occurs in basal cells that produce new cells instead of dying old cells in the skin. It is the most common form of skin cancer. It is more common in areas exposed to prolonged sunlight, such as the head and neck. It may appear as pink growths, persistent wounds, red spots on the skin. Such lesions grow slowly and rarely metastasize. If diagnosed early, treatment becomes easier [52].
- Benign keratosis-like lesions (bkl): In the dataset, three different types of lesions with no cancer characteristics were grouped under this categorical type. Types of lesions in this group are seborrheic keratosis, solar lentigo, and lichenoid keratosis. Seborrheic keratosis, which is harmless and does not require treatment, is brown, black, or light skin-colored and is usually found

on the back, neck, head, and chest. Solar lentigo, which is especially common in people over the age of 40, is a harmless, dark skin spot caused by exposure to UV radiation. Lichenoid keratosis or Lichen planus-like keratosis are harmless lesions that are usually seen between the ages of 50 and 70 and can develop in less than 3 months [53].

- **Dermatofibroma (df):** They are small, non-cancerous skin growths that usually appear on the lower leg, upper arm, and upper back. It can be of different colors and their colors may change over time. It is more common in adults than children and women than men. It is more common in people with a weakened immune system [54].
- **Melanocytic nevi (nv):** They are neoplasms caused by the proliferation of melanocytes, which are pigment-producing cells in the skin. It usually occurs in early childhood due to ultraviolet rays from the sun. Melanocytic nevi are more symmetrical in structure, unlike melanoma. They are benign skin lesions that appear as moles in various parts of the body [55].
- **Melanoma (mel):** It is the most dangerous type of skin cancer that occurs in melanocyte cells. Although the reason for its occurrence is not clear, it is caused by exposure to ultraviolet radiation from the sun or artificial light sources. It is mostly seen in sun-exposed areas of the body such as the face, back, arms, neck, and legs. It is rarely seen in the inner parts of the body such as the nose and throat. The risk of melanoma, which has an increased prevalence in the world, is higher in people under 40 and women. Moles with asymmetric shape, irregular edges, different colors, over 6 mm and tend to grow may be signs of melanoma. If left untreated, it can spread to different parts of the body and cause death. In the case of early diagnosis, it can be treated with surgical operation [56].
- **Vascular lesions (vasc):** They are lesions of different appearance, usually seen in the head and neck, occur on the skin and underlying tissues, and are generally known as birthmarks. Most of them are congenital but rarely can occur later and are rarely malignant [57].

The dataset includes 6705 melanocytic nevi, 1113 melanomas, 1099 benign keratoses, 514 basal cell carcinomas, 327 actinic keratoses, 142 vascular lesions, 115 dermatofibromas. Sample images of the seven different types mentioned are given in Fig. 2, and the graph showing the distribution of the lesion types is shown in Fig. 3.

2.2 Convolutional neural network (CNN)

CNN is a computer vision technique that automatically extracts and identifies features from images. As in many

other fields, it has gained popularity in medical image analysis due to its high performance. A typical CNN consists of convolution, activation, pooling, dropout, fully connected, and classification layers. The general structure of a CNN model is shown in Fig. 4.

The pixels of an image to be processed are CNN's inputs. In the convolution layer feature detector matrices, also known as filters, are stridden over the original input pixels to obtain a subset of features. Convolution is the main operation of CNN and provides automatic extraction of features. In the pooling stage, a dimension reduction operation is performed with filters applied to the input matrix. The filter window, also called the pool, is stridden over the original input matrix, and the reduction process is performed by taking the maximum, minimum, or average of the values in the pool.

In neural network models, especially if the dataset used for training is insufficient, it is possible to encounter overfitting problems. A dropout operation applied to overcome this problem, increases the success of the network to adapt to different situations by randomly disabling a certain ratio of neurons during the training process. The fully connected layer takes part in the transition to the classification stage. After feature extraction and pooling operations, the output matrix is flattened for transfer to the classifier.

3 Experimental study and findings

In this study, the HAM10000 dataset consisting of 10,015 dermoscopic images were classified into seven different types with the proposed CNN model. Before training the model, a series of pre-processing was performed on images. Data augmentation technique was used during training. The trained model was saved and connected to a web application via a REST-API. The application was tested by seven expert dermatologists with 5 to 9 years of professional experience. In the test process, the predictions made by the model and the diagnoses of the experts were compared. Experts gave feedback about the application's performance. During the testing process, a set of test images split from the existing dataset was used. Besides, the experts used their test images.

3.1 Pre-processing and data augmentation

The dataset consists of two basic parts. The first part is a metadata file containing descriptive information for lesion images. The metadata file includes the diagnosis of the lesion, the age, and gender of the patient, the location of the lesion, and the file path of the lesion image. The second and main part of the dataset consists of image files.

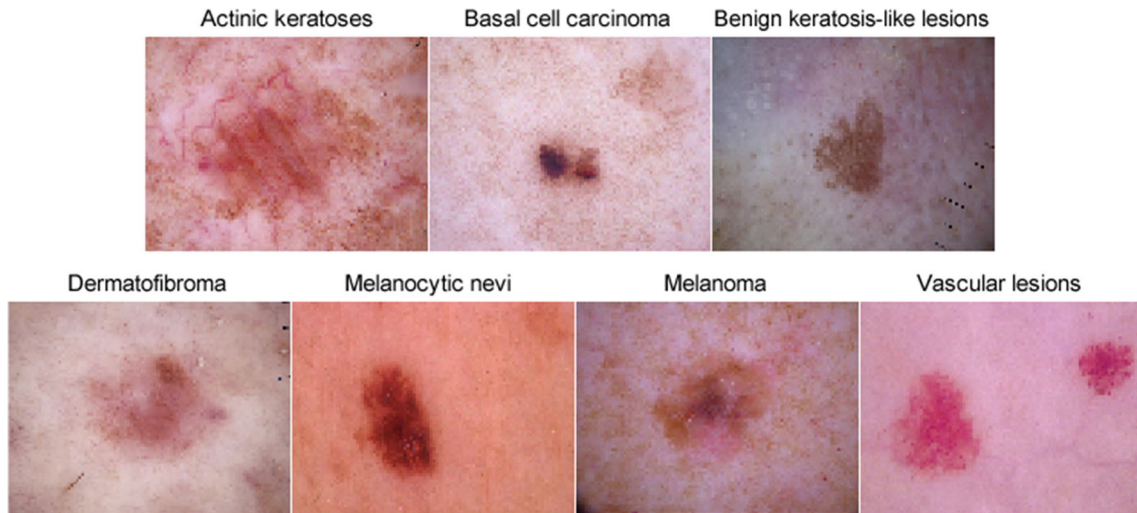


Fig. 2 Sample images of lesion types in the dataset

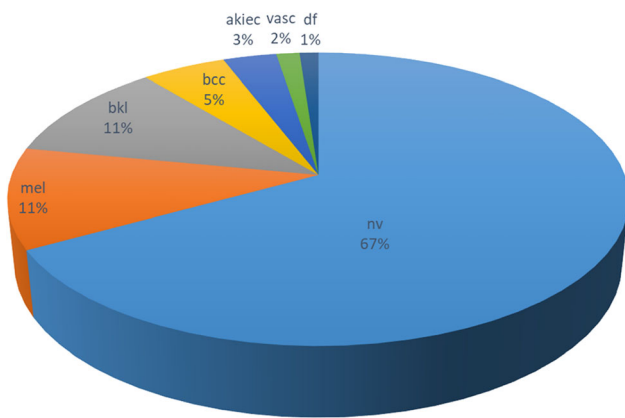


Fig. 3 Distribution of lesion types in the dataset

The purpose of this study is to make a diagnosis using images of the lesions only. Therefore, the metadata file was rearranged to contain only two columns, the image file path, and the lesion type. The labels of each lesion in text format were digitized to express a value between 0 and 6. The label codes for each type are listed in Table 1.

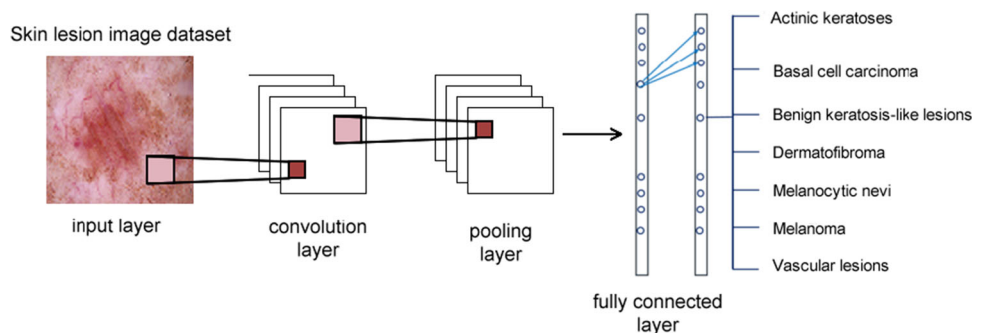
Table 1 Lesion labels

Lesion type	akiec	bcc	bkl	df	nv	mel	vasc
Encoded Label	0	1	2	3	4	5	6

The original images are 600×400 pixels and in RGB format. The larger the image size, the higher the processing load. Reducing the image size improves the processing speed. For this purpose, all images in the dataset are resized to 100×75 pixels. Since color has a distinctive feature in determining the lesion type, the original colors of the images were preserved. The contrast of each input image was increased and a sharpening filter was applied to make the lesion area more prominent.

The dataset was randomly divided into 80% training and 20% test set. There were 8012 images in the training set and 2003 images in the test set. The test set was reserved for the test to be carried out with the web application. 10% of the training set was randomly split to be used in the validation process during the training. There are 801

Fig. 4 The general structure of a CNN model



images in the validation set. There were images from each of the seven different lesion classes in the train, test, and validation sets.

Working with a relatively small amount of data causes an overfitting problem. Although there are more than 10 thousand images in the dataset used, the number of images in some of the seven different classes is small. Enriching the dataset is important to increase prediction success. It may be troublesome to obtain original data, especially in such studies with images related to human health. Therefore, programmatic data augmentation is a technique frequently used in computer image analysis applications performed with CNN.

In the data augmentation process, new images are produced using techniques such as rotation, shifting, reflecting, scaling and flipping. The data augmentation increases the model's ability to adapt to different situations. In this study, a data augmentation was performed on random images selected from the original dataset by applying 10 degrees of rotation, 10% zoom, horizontal and vertical shifting of 10%. These ratios are experimental and can be tuned, but these ratios provided a good result for this study.

3.2 Building and training the CNN model

The general structure of the proposed CNN model for skin lesion classification is shown in Fig. 5.

The model can be handled as 3 separate blocks. 3-channel, 75×100 pixels sized images are the inputs of the model. Convolution operations are performed on the first two layers of the first block. In each of these layers, there are 32 filters of 3×3 size. After convolution processes, maximum pooling is performed with a pool size of 2×2 . Then 25% dropout is applied.

The second block has the same order of operation as the first, but there are parametric differences. In the first two layers of the second block, convolution operations are performed with 64 filters with 3×3 dimensions. Then, maximum pooling is applied with a 2×2 pool size and then a 40% dropout is applied.

The ReLU activation function and the same padding technique are used in all convolution layers of the first two blocks. The output matrix obtained at the end of the second block is flattened and transferred to the third block.

In the third block, there is a neural network that will carry out the classification process. The first layer of this block consists of 256 neurons, the second layer consists of 128 neurons, and both use the ReLU as an activation function. After these two layers, a 50% dropout is applied. The last layer consists of 7 output neurons as many as the number of classes in the dataset. In this layer, the Softmax activation function is used for multi-class classification.

The layer structure of the proposed CNN model, the number of components in each layer, and the number of inputs of the layers are shown in Fig. 6. Besides, the summary of the developed model is given in Table 2.

In a CNN, only Convolutional and Fully Connected Layers have tunable parameters. Accordingly, the proposed CNN model has 7 weighted layers, 6 hidden layers, and 1 output layer in addition to an input layer. As details shown in Table 2, the first two convolution layers each have 32 neurons, and the third and fourth convolution layers each have 64 neurons. In the neural network in the last section where the classification takes place, there are two hidden layers with 256 and 128 neurons and an output layer with 7 neurons. The model has a total of 7.472.423 trainable parameters.

The hyperparameters used for the training of the proposed CNN model and their values are given in Table 3.

3.3 Performance Measurements

There are several metrics used to evaluate the performance of a machine learning model. The accuracy metric is used to evaluate the overall success of the model (Eq. 2).

$$\text{accuracy} = \frac{\text{(number of correctly classified samples)}}{\text{(total number of samples)}} \quad (2)$$

The model was run 100 epochs with GPU acceleration. In the trials for training the model, after 100 epochs there

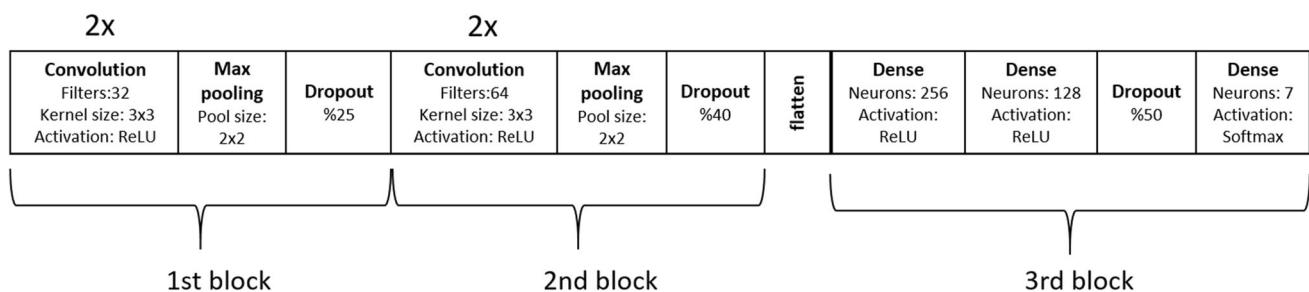


Fig. 5 The general structure of the proposed model

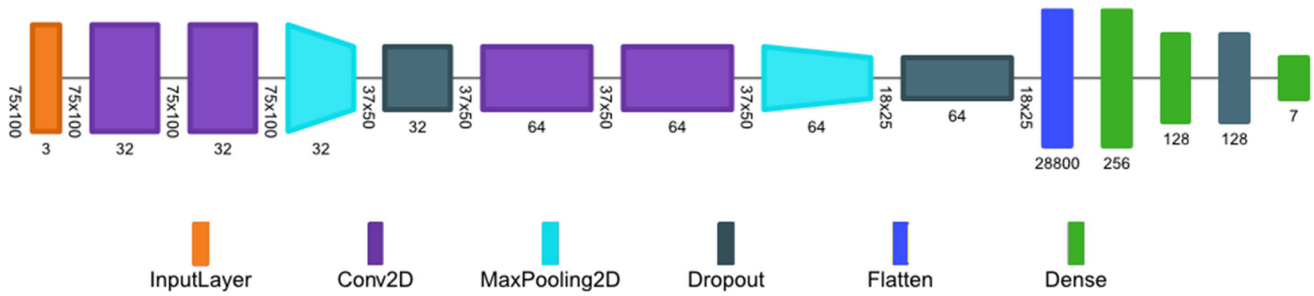


Fig. 6 The layer structure of the CNN model

Table 2 Summary of the model

Layer (type)	Output Shape	Parameters
Convolution 2D	(None, 75, 100, 32)	896
Convolution 2D	(None, 75, 100, 32)	9248
Max Pooling 2D	(None, 37, 50, 32)	0
Dropout	(None, 37, 50, 32)	0
Convolution 2D	(None, 37, 50, 64)	18,496
Convolution 2D	(None, 37, 50, 64)	36,928
Max Pooling 2D	(None, 18, 25, 64)	0
Dropout	(None, 18, 25, 64)	0
Flatten	(None, 28,800)	0
Dense	(None, 256)	7,373,056
Dense	(None, 128)	32,896
Dropout	(None, 128)	0
Dense	(None, 7)	903

Table 3 Model training parameters

Parameter	Value
Batch size	10
Number of epochs	100
Number of iterations	721
Optimizer	Adam
Optimizer parameters	lr = 0.0001, beta1 = 0.9, beta2 = 0.999

was no significant increase in the accuracy of the model and the accuracy curve began to oscillate in a similar range. The situation was similar for the loss values. For this reason, considering that increasing the number of epochs does not increase the performance while extending the training period, it was deemed appropriate to limit the training process to 100 epochs.

The accuracy change graph obtained by running the model for 100 epochs is given in Fig. 7. The highest accuracy value was obtained as 91.51%.

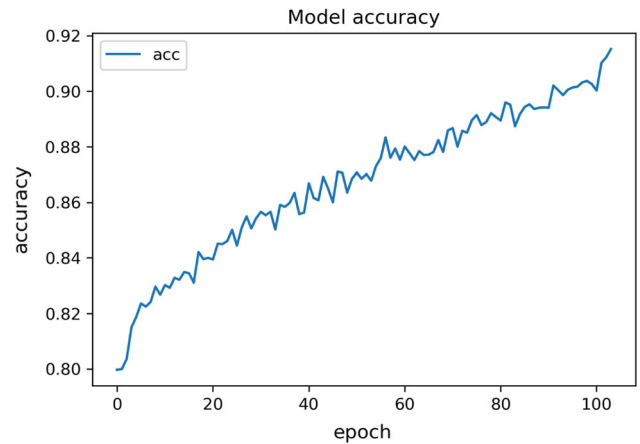


Fig. 7 The accuracy graph of the model

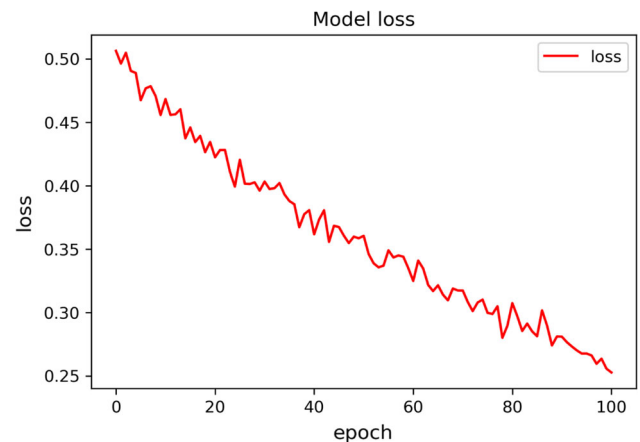


Fig. 8 The loss graph of the model

The decrease in the loss value in parallel with the increase in the accuracy of the model is shown in Fig. 8.

Accuracy is an effective parameter to characterize performance if the dataset is balanced. However, when working with imbalanced datasets, the system should be evaluated with more performance metrics. At this stage, confusion matrices are used. The confusion matrix provides detailed information about correct and incorrect classifications (Fig. 9).

		Predicted class	
		+	-
Actual class	+	TP True Positives	FN False Negatives
	-	FP False Positives	TN True Negatives

Fig. 9 Confusion matrix

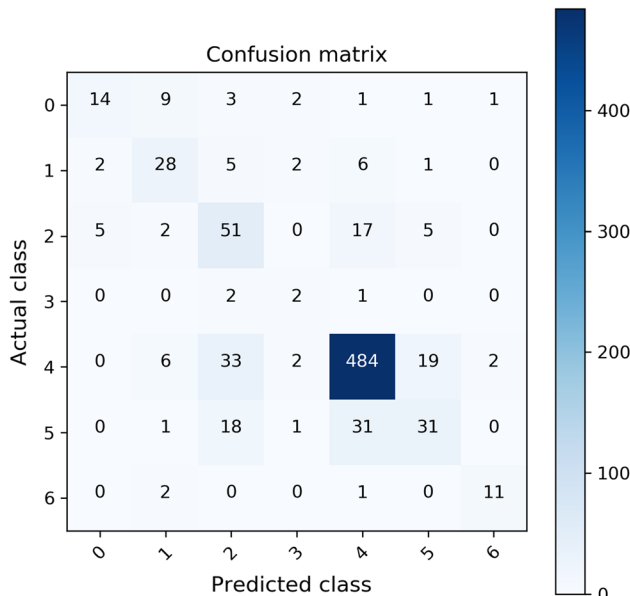


Fig. 10 The confusion matrix of the model

As a result of training the model, the confusion matrix obtained for seven different classes is shown in Fig. 10. With the help of the confusion matrix, it can be seen in more detail which classes are distinguished and which are confused.

Different metrics are calculated to express the performance of the classifier with the parameters obtained from the confusion matrix. These metrics are explained in Table 4.

The Precision, Recall, and F1 scores calculated for each of the seven classes are given in Table 5.

When the confusion matrix and performance scores are examined, it is seen that the model has the highest success in classifying Melanocytic nevi (nv). The lowest success is seen in the classification of Dermatofibromas (df). The

Table 5 Performance scores of the model

Lesion Class	Precision	Recall	F1 Score
0—akiec	0.67	0.45	0.54
1—bcc	0.58	0.64	0.61
2—bkl	0.45	0.66	0.54
3—df	0.30	0.41	0.35
4—nv	0.91	0.90	0.90
5—mel	0.58	0.38	0.45
6—vasc	0.79	0.79	0.79

probable reason for this low success is that the number of samples of this type in the dataset is around 2%. The most confusing lesion types were melanoma and melanocytic nevi.

According to test results, the comparison of this study with similar recent studies is given in Table 6. Most of the studies listed in Table 6 used the data sets that served as the training sets for different ISIC challenges. The HAM10000 dataset used in this study was presented in the ISIC 2018 challenge. It contains 10,015 images of 7 different types. ISIC 2017 dataset is smaller than HAM10000 and consists of 2000 images of 3 different types.

Salian et al. [39], Pereira et al. [22], Bi et al. [31], Serte, and Demirel [40] conducted their studies on limited datasets. For this reason, the generalizability of the results was limited compared to this study. Other studies in Table 5 were conducted on the same dataset used in this study. Milton [42], Zhang et al.[30], Aldwgeri and Abubacker [43], Liu et al. [41], Almaraz-Damian et al. [23], Mahbood et al. [35], Yilmaz and Trocan [37] used the transfer learning method and reported the success of different pre-trained models. In the transfer learning approach, since it is the case of using the trained parameters of a network that has previously been trained on a very large amount of data, it is possible to obtain high accuracy for different problems. The highest accuracy they reached with this method was 92%. The model proposed in this study was created and trained from scratch. Although it was trained with fewer data and in limited time compared to pre-trained models, it performed well to compete with these studies. Almaraz-Damian et al. achieved 92% accuracy by using

Table 4 The performance metrics

Metric	Formula	Explanation
Precision	$TP / (TP + FP)$	How accurate the positive predictions are
Recall (sensitivity)	$TP / (TP + FN)$	Coverage of actual positive sample
F1 Score	$2 * (Precision * Recall) / (Precision + Recall)$	The harmonic mean of precision and recall

Table 6 Comparison of the results obtained in this study with the literature

Reference	Dataset	Method	The highest accuracy (%)
Serte and Demirel [40]	ISIC 2017 dataset	AlexNet and ResNet-18	81–83
Milton [42]	HAM1000 dataset	PNASNet-5-Large	76
Zhang et al. [30]	ISIC 2017 dataset	ARL-CNN	85
Aldwgeri and Abubacker [43]	HAM1000 dataset	VGG-Net	77
Salian et al. [39]	HAM1000 dataset	Pre-trained and custom CNN model	80
Liu et al. [41]	ISIC 2017 dataset	DenseNet and ResNet	87
Almaraz-Damian et al. [23]	HAM10000 dataset	Using handcrafted features, with Linear Regression and Support Vector Machines	92
Mahbood et al. [35]	HAM10000 dataset	EfficientNet and SeReNeXt-50	86
Yilmaz and Trocan [37]	ISIC dataset	AlexNet, GoogLeNet, ResNet50	91–92
Bi et al. [31]	7-PC dataset contains 1011 samples	HcCNN	74
Pereira et al. [22]	Dermofit dataset contains 1300 images	SVM and feed-forward network	77–78
Proposed CNN model	HAM1000 dataset	Custom CNN architecture	91.51

handcrafted features. In this study, a similar level of accuracy was obtained with automatic feature extraction without the need for manual processing that requires special expertise.

3.4 Web application and expert evaluation

The developed classification model was saved for use in a web application infrastructure. A REST API (Representational State Transfer Application Programming Interface) was built on the model using the Python Flask framework. The developed web application communicates with the classification model via this API.

The web application has a very simple interface for experts to easily use. The application is responsive to work smoothly on different devices and resolutions. For classification, a specific test image can be uploaded to the system (Select Image) or a random image can be selected from the test set (Get Random). The predictions are displayed on the page after the selected images were classified by the model (Fig. 11).

The web application developed for skin lesion classification was tested by seven expert dermatologists who voluntarily participated in this study. The professional experience of these dermatologists varies from 5 to 9 years. The test of the application was carried out in two phases. In the first phase, the skin lesions were first classified by the model and then confirmed by the expert dermatologists. In this way, the success of the application in predicting the skin lesion type correctly was tested in practice. In the

**Fig. 11** The skin lesion classifier web application

second phase, the skin lesions were first diagnosed by the expert dermatologists and then classified by the model. At this phase, the corrective role of the application against possible misdiagnoses was tested.

In two phases and, for seven expert dermatologists, 50 images at each, a total of 700 images were used for testing. The experts can take images randomly from the test set, as well as upload images from their archives. In the first phase, the selected image was first classified by the model and then the expert checked the result. According to the feedback received, the model correctly classified 316 of the 350 images used for testing in the first phase. Details of the test results obtained for each expert in the first phase are given in Table 7.

Table 7 First phase test results

Expert	Number of correct predictions	Number of wrong predictions	Success rate (%)	Most confused lesion types
1	46	4	92	nv-mel
2	44	6	88	nv-mel, bcc- akiec
3	46	4	92	bkl-nv, nv-mel
4	45	5	90	nv-mel
5	45	5	90	nv-mel, bkl-akiec
6	44	6	88	bkl-nv, nv-mel
7	46	4	92	nv-mel
	316	34	90,285	

Considering these results, the test success of the model in the first phase was found as 90.28%. This result coincided with the model's training accuracy. In the tests, it was observed that the most confusing lesion types were Melanocytic nevi and Melanoma. Besides, it was observed that benign keratosis-like lesions were confused with Melanocytic and Actinic keratosis and Basal cell carcinomas with Actinic keratosis. According to the test images used, it was seen that the classification success was reduced in cases such as hair growth on the lesion and dark skin color. As a result, it was concluded that it would be beneficial to make corrections on the images before analysis in future works.

In the second phase, the lesion images taken randomly from the test dataset were first diagnosed by the expert dermatologists. Then the same lesion images were classified by the model. In this phase, experts misdiagnosed 39 of the 350 images, and the predictions made by the model enabled the correct diagnosis to be made. According to these results, the model played a corrective role by 11.14%. In the literature, studies are reporting that machine learning techniques are more successful than humans in classifying skin lesion images [58]. In this study, instead of comparing the model developed with the experts, the success of the model in providing support to the experts is emphasized.

Changes in the success of the model are possible by increasing the number of tests and the number of images used for testing. In future works, increasing the number of dermatologists participating in the test will also provide more generalizable results.

4 Conclusion

Pigmented skin lesions are patches that originate from melanocyte cells. The uncontrolled division of melanocyte cells causes skin cancers such as melanoma and can have fatal consequences. Dermoscopy is widely used in the diagnosis of skin lesions. The diagnosis of dermoscopic

images is made by expert dermatologists. The use of computer-aided systems comes to the fore for the insufficient number of trained experts and for minimizing human-induced errors. CNN, a deep learning technique that automatically extracts the features within patterns, produces successful results in the field of computer vision. In this study, using the HAM10000 dataset, images of skin lesions were classified into seven different types with a CNN model developed. The model reached an accuracy of 91.51%. A web application was built on the developed model and a two-stage test was carried out with seven expert dermatologists. In the first phase, the predictions made by the model were then verified by the experts, and an accuracy of 90.28% was achieved. In the second phase, 11.14% of the misdiagnoses among the diagnoses made by the experts were corrected by the model. According to the test scores and the experts' feedback, it was concluded that the developed model can be used as a decision support system. However, it will be useful to test the model with more images and more experts to get generalizable results.

In future works, it is aimed to achieve higher success by improving the model and enriching the dataset. Studies on lesion diagnosis in real-time images are also planned. Besides, it is aimed to reach generalizable results by testing the model with more experts and to make the application usable in practice.

Declarations

Conflict of interest The author has no conflicts of interest to declare that are relevant to the content of this article.

References

- Gallagher RP, Lee TK, Bajdik CD, Borugian M (2010) Ultraviolet radiation. Chronic diseases and injuries in Canada, vol. 29
- Saladi RN, Persaud AN (2005) The causes of skin cancer: a comprehensive review. *Drugs Today* 41(1):37–54

3. Guy GP Jr, Thomas CC, Thompson T, Watson M, Massetti GM, Richardson LC (2015) Vital signs: melanoma incidence and mortality trends and projections—United States, 1982–2030. *MMWR Morb Mortal Wkly Rep* 64(21):591
4. Okur E, Turkan M (2018) A survey on automated melanoma detection. *Eng Appl Artif Intell* 73:50–67
5. WHO Ultraviolet (UV) radiation and skin cancer (2020) [https://www.who.int/news-room/q-a-detail/ultraviolet-\(uv\)-radiation-and-skin-cancer](https://www.who.int/news-room/q-a-detail/ultraviolet-(uv)-radiation-and-skin-cancer). Accessed Aug. 20, 2020
6. Türk Dermatoloji Derneği (2020) http://turkdermatoloji.org.tr/haber_arsivi/detay/17. Accessed Aug. 20, 2020
7. Kittler H, Pehamberger H, Wolff K, Binder M (2002) Diagnostic accuracy of dermoscopy. *Lancet Oncol* 3:159–165. [https://doi.org/10.1016/S1470-2045\(02\)00679-4](https://doi.org/10.1016/S1470-2045(02)00679-4)
8. Carli P et al (2003) Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. *Br J Dermatol* 148:981–984. <https://doi.org/10.1046/j.1365-2133.2003.05023.x>
9. Masood A, Al-Jumaily A (2013) Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms. *Int J Biomed Imaging* 2013:323268. <https://doi.org/10.1155/2013/323268>
10. Nanni L, Lumini A, Ghidoni S (2018) Ensemble of deep learned features for melanoma classification. *arXiv preprint arXiv:1807.08008*
11. Celebi ME et al (2007) A methodological approach to the classification of dermoscopy images. *Comput Med Imaging Graph* 31(6):362–373
12. Oliveira RB, Mercedes Filho E, Ma Z, Papa JP, Pereira AS, Tavares JMR (2016) Computational methods for the image segmentation of pigmented skin lesions: a review. *Comput Methods Progr Biomed* 131:127–141
13. Barata C, Celebi ME, Marques JS (2014) Improving dermoscopy image classification using color constancy. *IEEE J Biomed Health Informat* 19(3):1146–1152. <https://doi.org/10.1109/JBHI.2014.2336473>
14. Schaefer G, Rajab M, Celebi ME, Iyatomi H (Oct. 2010) Colour and contrast enhancement for improved skin lesion segmentation. *Comput Med Imag Soc* 35:99–104. <https://doi.org/10.1016/j.compmedimag.2010.08.004>
15. Iyatomi H, Celebi ME, Schaefer G, Tanaka M (Oct. 2010) Automated color calibration method for dermoscopy images. *Comput Med Imag Soc* 35:89–98. <https://doi.org/10.1016/j.compmedimag.2010.08.003>
16. Zhang J, Xie Y, Wu Q, Xia Y (2019) Medical image classification using synergic deep learning. *Med Imag Anal* 54:10–19. <https://doi.org/10.1016/j.media.2019.02.010>
17. Fatima R, Khan MZA, Dhruve KP (2012) Computer aided multi-parameter extraction system to aid early detection of skin cancer melanoma. *Int J Comput Sci Netw Secur* 12(10):74–86
18. Qian C et al. (2018) A detection and segmentation architecture for skin lesion segmentation on dermoscopy images. *ArXiv*, [arXiv:1809.03917](https://arxiv.org/abs/1809.03917)
19. Ercal F, Chawla A, Stoecker WV, Lee HC, Moss RH (1994) Neural network diagnosis of malignant melanoma from color images. *IEEE Trans Biomed Eng* 41(9):837–845
20. Erçal F, Lee HC, Stoecker WV, Moss RH (1994) Skin cancer diagnosis using hierarchical neural networks and fuzzy systems.
21. Binder M, Steiner A, Schwarz M, Knollmayer S, Wolff K, Pehamberger H (1994) Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. *Br J Dermatol* 130(4):460–465
22. Pereira PMM et al (2020) Skin lesion classification enhancement using border-line features—the melanoma vs nevus problem. *Biomed Signal Process Control* 57:101765. <https://doi.org/10.1016/j.bspc.2019.101765>
23. Almaraz-Damian J-A, Ponomaryov V, Sadovnychiy S, Castillejos-Fernandez H (2020) Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. *Entropy* 22(4):484
24. Kavitha JC, Suruliandi A (2016) Texture and color feature extraction for classification of melanoma using SVM. pp. 1–6
25. Hu Z, Tang J, Wang Z, Zhang K, Zhang L, Sun Q (2018) Deep learning for image-based cancer detection and diagnosis—a survey. *Pattern Recogn* 83:134–149. <https://doi.org/10.1016/j.patcog.2018.05.014>
26. Nasiri S, Jung M, Helsper J, Fathi M (2018) Deep-CLASS at ISIC machine learning challenge
27. Naylor P, Laé M, Reyal F, Walter T (2017) Nuclei segmentation in histopathology images using deep neural networks. In: 2017 IEEE 14th International symposium on biomedical imaging (ISBI 2017), pp. 933–936
28. Mahbod A, Schaefer G, Ellinger I, Ecker R, Pitiot A, Wang C (2019) Fusing fine-tuned deep features for skin lesion classification. *Comput Med Imaging Graph* 71:19–29. <https://doi.org/10.1016/j.compmedimag.2018.10.007>
29. Kawahara J, BenTaieb A, Hamarneh G (2016) Deep features to classify skin lesions. In: 2016 IEEE 13th International symposium on biomedical imaging (ISBI), pp. 1397–1400
30. Zhang J, Xie Y, Xia Y, Shen C (2019) Attention residual learning for skin lesion classification. *IEEE Trans Med Imag* 38(9):2092–2103. <https://doi.org/10.1109/TMI.2019.2893944>
31. Bi L, Feng DD, Fulham M, Kim J (2020) Multi-Label classification of multi-modality skin lesion via hyper-connected convolutional neural network. *Pattern Recogn* 107:107502. <https://doi.org/10.1016/j.patcog.2020.107502>
32. Ayan E Ünver HM (2018) Data augmentation importance for classification of skin lesions via deep learning. In: 2018 Electric electronics, computer science, biomedical engineering's meeting (EBBT), pp. 1–4.
33. Menegola A, Fornaciali M, Pires R, Bittencourt FV, Avila S, Valle E (2017) Knowledge transfer for melanoma screening with deep learning. In: 2017 IEEE 14th International symposium on biomedical imaging (ISBI 2017), pp. 297–300
34. Lopez AR, Giro-i-Nieto X, Burdick J, Marques O (2017) Skin lesion classification from dermoscopic images using deep learning techniques. In: 2017 13th IASTED International conference on biomedical engineering (BioMed), pp. 49–54
35. Mahbod A, Schaefer G, Wang C, Dorffner G, Ecker R, Ellinger I (2020) Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Comput Methods Programs Biomed* 193:105475. <https://doi.org/10.1016/j.cmpb.2020.105475>
36. Gessert N, Nielsen M, Shaikh M, Werner R, Schlaefer A (2020) Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX* 7:100864. <https://doi.org/10.1016/j.mex.2020.100864>
37. Yilmaz E, Trocan M (2020) Benign and malignant skin lesion classification comparison for three deep-learning architectures. In: Asian conference on intelligent information and database systems, pp. 514–524.
38. Li Y, Shen L (2018) Skin lesion analysis towards melanoma detection using deep learning network. *Sensors* 18(2):556. <https://doi.org/10.3390/s18020556>
39. Salian AC, Vaze S, Singh P, Shaikh GN, Chapaneri S, Jayaswal D (2020) Skin lesion classification using deep learning architectures. In: 2020 3rd International conference on communication system, computing and IT applications (CSCITA), 2020, pp. 168–173, doi: <https://doi.org/10.1109/CSCITA47329.2020.9137810>

40. Serte S, Demirel H (2019) Gabor wavelet-based deep learning for skin lesion classification. *Comput Biol Med* 113:103423. <https://doi.org/10.1016/j.combiomed.2019.103423>
41. Liu L, Mou L, Zhu XX, Mandal M (2020) Automatic skin lesion classification based on mid-level feature learning. *Comput Med Imag Graph* 84:101765. <https://doi.org/10.1016/j.compmedimag.2020.101765>
42. Milton MAA (2019) Automated skin lesion classification using ensemble of deep neural networks in ISIC 2018: Skin lesion analysis towards melanoma detection challenge
43. Aldwgeri A, Abubacker NF (2019) Ensemble of deep convolutional neural network for skin lesion classification in dermoscopy images. In: *International visual informatics conference*, pp. 214–226.
44. Anjum MA, Amin J, Sharif M, Khan HU, Malik MSA, Kadry S (2020) Deep semantic segmentation and multi-class skin lesion classification based on convolutional neural network. *IEEE Access* 8:129668–129678. <https://doi.org/10.1109/ACCESS.2020.3009276>
45. Gessert N et al. (2018) Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting. *CoRR*, vol. abs/1808.01694, 2018, [Online]. Available: <http://arxiv.org/abs/1808.01694>
46. Rezvantalab A, Safigholi H, Karimijeshni S (2018) Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. *CoRR*, vol. abs/1810.10348, 2018, [Online]. Available: <http://arxiv.org/abs/1810.10348>
47. Goyal M, Rajapakse JC (2018) Deep neural network ensemble by data augmentation and bagging for skin lesion classification. *CoRR*, vol. abs/1807.05496, 2018, [Online]. Available: <http://arxiv.org/abs/1807.05496>
48. Kassani SH, Kassani PH (2019) A comparative study of deep learning architectures on melanoma detection. *Tissue Cell* 58:76–83. <https://doi.org/10.1016/j.tice.2019.04.009>
49. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci data* 5:180161
50. ISIC Archive (2020) <https://www.isic-archive.com>. Accessed Aug. 22, 2020
51. Callen JP, Bickers DR, Moy RL (1997) Actinic keratoses. *J Am Acad Dermatol* 36(4):650–653
52. Rubin AI, Chen EH, Ratner D (2005) Basal-cell carcinoma. *N Engl J Med* 353(21):2262–2269
53. Nagrani N, Jaimes N, Oliviero MC, Rabinovitz HS (Jul. 2018) Lichen planus-like keratosis: clinical applicability of in vivo reflectance confocal microscopy for an indeterminate cutaneous lesion. *Dermatol Pract Concept* 8(3):180–183. <https://doi.org/10.5826/dpc.0803a06>
54. Alves JVP, Matos DM, Barreiros HF, Bártole EAFLF (2014) Variants of dermatofibroma—a histopathological study. *An Bras Dermatol* 89(3):472–477. <https://doi.org/10.1590/abd1806-4841.20142629>
55. Damsky WE, Bosenberg M (Oct. 2017) Melanocytic nevi and melanoma: unraveling a complex relationship. *Oncogene* 36(42):5771–5792. <https://doi.org/10.1038/onc.2017.189>
56. Usher-Smith JA, Emery J, Kassianos AP, Walter FM (2014) Risk prediction models for melanoma: a systematic review. *Cancer Epidemiol Prev Biomark* 23(8):1450–1463
57. Brahmabhatt AN, Skalski KA, Bhatt AA (Feb. 2020) Vascular lesions of the head and neck: an update on classification and imaging review. *Insights Imag* 11(1):19. <https://doi.org/10.1186/s13244-019-0818-3>
58. Tschandl P et al (2019) 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(7):938–947. [https://doi.org/10.1016/S1470-2045\(19\)30333-X](https://doi.org/10.1016/S1470-2045(19)30333-X)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.