Original Paper

Skin Diseases Classification Using Deep Leaning Methods

ANCA-LOREDANA UDRIȘTOIU[1], ARIANA ELENA STANCA[1], ALICE ELENA GHENEA[2], CORINA MARIA VASILE[2], MIHAELA POPESCU[2], ȘTEFAN CRISTINEL UDRIȘTOIU[1], ANDREEA VALENTINA IACOB[1], STEFAN CASTRAVETE[4], LUCIAN GHEORGHE GRUIONU[3], GABRIEL GRUIONU[3]

[1]*Faculty of Automation, Computers and Electronics, University of Craiova, Craiova, Romania*

[2]*Faculty of Medicine and Pharmacy Craiova of Craiova, Romania*

[3]*Faculty of Mechanics, University of Craiova, Craiova, Romania*

[4]*Caelynx Europe LLC, Craiova, Romania*

ABSTRACT: Due to the high incidence of skin tumors, the development of computer aided-diagnosis methods will become a very powerful diagnosis tool for dermatologists. The skin diseases are initially diagnosed visually, through clinical screening and followed in some cases by dermoscopic analysis, biopsy and histopathological examination. Automatic classification of dermatoscopic images is a challenge due to fine-grained variations in lesions. The convolutional neural network (CNN), one of the most powerful deep learning techniques proved to be superior to traditional algorithms. These networks provide the flexibility of extracting discriminatory features from images that preserve the spatial structure and could be developed for region recognition and medical image classification. In this paper we proposed an architecture of CNN to classify skin lesions using only image pixels and diagnosis labels as inputs. We trained and validated the CNN model using a public dataset of 10015 images consisting of 7 types of skin lesions: actinic keratoses and intraepithelial carcinoma/Bowen disease (akiec), basal cell carcinoma (bcc), benign lesions of the keratosis type (solar lentigine/seborrheic keratoses and lichen-planus like keratosis, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhages, vasc).

KEYWORDS: *Machine learning, deep learning, convolutional neural network, dermatoscopic images, medical.*

Introduction

Early detection of skin cancers, including melanoma and non-melanoma skin cancers is crucial.

Skin cancer is growing globally, for example in the United States, the number of new cases of skin melanoma was 22.2 per 100,000 men and women per year.

The number of deaths was 2.5 per 100,000 men and women per year, based on 2012-2016 cases and deaths [1, 2].

Approximately 2.3 percent of men and women will be diagnosed with skin melanoma at some point during their lifetime, based on 2014-2016 data [2-4].

In 2016, almost 1,195,608 people have skin melanoma in the United States [2].

Melanoma has been reported as the 6th most common of all cancer cases [5].

On the other hand, most non-melanoma skin cancers, which are responsible for 4.3-5.4 million new cases each year in the United States [3,6], can be treated simply by surgical removal.

Early detection of all skin cancers is very important to prevent the progression to advanced stages.

Even if the percentage of patients with malignant skin tumors is not that high, the primary screening control of patients is important and determine which patients are at risk and should be transferred to the dermatologists [6-8].

Thus, any method that can accurately give the probability of malignancy by analyzing a simple image of the tumor would be very helpful for primary diagnosis [5].

In this context, the development of artificial intelligence (AI) and computer-aided systems that can quickly classify skin tumor images, similar to trained dermatologists, is the best solution.

Previous dermatological computer-aided systems used to classify dermatoscopic images have missed the generalization capability due to insufficient data.

Therefore, the labelling of dermatoscopic images is essential to develop diagnostic images methods together to powerful deep learning techniques.

10.12865/CHSJ.46.02.06

The purpose of our study was to assess the accuracy of CNNs for detection of dermatological lesions which are diagnosed as actinic keratoses and intraepithelial carcinoma/ Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc) [1].

Many previous methods, which have used traditional machine learning algorithms, require image preprocessing, segmentation, and visual function extraction before classification.

Our method doesn't use hand-crafted features; the algorithm is trained only from image labels and raw pixels of dermoscopic images.

Methods and Materials

Study description

In this study we used the HAM10000 ("Human against Machine with 10000 training images") dataset [1].

Dermatoscope images were collected from different populations, acquired and stored in different ways.

The original dataset consists of 10015 dermatoscopic images consisting of all important diagnostic categories in the realm of pigmented lesions: actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc) [1].

The diagnosis of images was collected through different methods: more than 50% of lesions are confirmed by histopathology, the other cases were confirmed by either follow-up examination, expert consensus, or confirmation by in-vivo confocal microscopy [1].

To overcome the problems caused by having an imbalanced dataset, we used data augmentation.

This technique generates more training data from existing images by augmenting samples through a number of random transformations.

The random transformations used on images were: *rotation_range* randomly rotates images, *width_shift* and *height_shift* randomly translate images vertically or horizontally, *shear_range* randomly applies shearing transformations and *zoom_range* randomly *zoomes* inside images [9].

The augmented dataset is a 7-class classification problem having 24,267 dermatoscopic images with an initial resolution of 600 x 450 pixels.

Of the total set of 24,267 dermatoscopic images, 23 018 images were used for training and 1,249 images for testing.

The split dataset could be observed in Table 1.

Table 1. Datasets for training and testing.

Diagnosis	Original Dataset	Augmented Datasets	
		Train	Test
AKIEC	327	2816	176
BCC (class	414	2999	187
BKL (class	1099	3123	187
DF (class 3)	165	2920	115
MEL(class	1113	3161	189
NV (class 5)	6705	5565	253
VASC (class	192	2434	142

CNN model

The proposed 4-CNN model

We proposed a four-layer convolutive model (4-CNN) to diagnose the dermatoscopic images. The relevant hyper-parameters within the deep learning framework [9,11] are defined as in Table 2.

The initial size of the dermatoscopic images was too large (600x450), bringing great difficulty to the algorithm performance. Therefore, the images size was primarily converted to 64 x 64.

Our proposed 4-CNN architecture contains 4 convolutional layers (Conv1, Conv2, Conv3, Conv4), 2 max-pooling layers (Pool1, Pool2), 6 batch normalization layers, 4 dropout layers, and 4 fully connected layers (FC1, FC2, FC3, FC4) as in Table 2.

Table 2. The parameters of the CNN model.

Layers	Conv1	Conv2	Pool1	Conv3	Conv4	Pool2	FC1	FC2	FC3	FC4
Kernel	3*3	3*3	2*2	3*3	3*3	2*2	-	-	-	-
Channel	32	32	32	64	64	64	512	128	50	7

Software and hardware

The model was written in Python (ver. 3.6.10, 64 bits) utilizing the open-source Keras (ver. 2.2.4) library [9] and the open-source TensorFlow (ver. 1.15.0) library [11] as backend for CNN models together with other scientific computing libraries as numpy [12] and scikit-learn [13].

The architecture on which we run the experiments is: Intel(R) Xeon(R) CPU E5-1620 v3 @ 3.50GHz, 32 GB RAM. We used a single NVIDIA Quadro K4200 GPU.

Experiments and results

The performance of our 4-CNN model is assessed taking into consideration that we are dealing with an unbalanced dataset.

Therefore, we use different metrics widely used in the field of medical diagnosis: Pre (precision), Se (sensitivity), Sp (specificity), test accuracy and area under the curve (AUC).

In order to evaluate the performance of each diagnosis, we established the current diagnosis as positive and the remaining ones as negative.

The average value of each evaluation metric in all diagnosis are computed, where the true positive refers to the set of images who belong to the positive class and are correctly classified; the false negative refers to the set of images who belong to the positive class but are misclassified as negative; the true negative are the set of patients who belong to the negative classes and are correctly classified and the false positives are patients' cases with negative classes that are predicted as current class.

Precision represents the fraction of cases classified as positive that belong to the positive class, as in (1).

$$\text{Pre} = \frac{|\text{true positives}|}{|\text{true positives}|+|\text{false positive}|} \quad (1)$$

Sensitivity represents how well the positive class was predicted, as in (2).

$$\text{Se} = \frac{|\text{true positive}|}{|\text{true positive}|+|\text{false negative}|} \quad (2)$$

Specificity is the complement to sensitivity and represents how well the negative class was predicted, as in (3).

$$\text{Sp} = \frac{|\text{true negative}|}{|\text{true negative}|+|\text{false positive}|} \quad (3)$$

Although widely used, the accuracy is inappropriate for unbalanced classification, as in (4).

$$\text{Accuracy} = \frac{|\text{correctly classified cases}|}{|\text{test cases}|} \quad (4)$$

The other two diagnostic tools that help interpret probabilistic prediction for multi-class classification are ROC Curves and Precision-Recall curves [14].

A ROC curve is a diagnostic diagram to summarize the model's behavior by calculating the true positive rate as in (5) and false positive rate as in (6) for a set of predictions of the model under different thresholds. The true positive rate is the recall or sensitivity.

A model that has no ability will be represented by a diagonal line from the bottom left to the top right. A good model will be a point in the top left of the plot.

$$\text{TruePositiveRate} = \frac{|\text{true positive}|}{|\text{true positive}|+|\text{false negative}|} \quad (5)$$

$$\text{FalsePositiveRate} = \frac{|\text{false positive}|}{|\text{false positive}|+|\text{true negative}|} \quad (6)$$

We calculated the micro and macro averaging to evaluate the overall performance of all diagnosis classes.

In micro averaging, we computed the performance from the individual true positives, true negatives, false positives, and false negatives of each diagnosis class. In macro averaging we computed the average performance of each diagnosis classes.

The Precision-Recall curves represents the balance between the true positive rate and the positive predictive value for a predictive model using different probability thresholds [15].

Results

Our 4-CNN method achieved a final accuracy of 93.6% and the surface below the curve of 0.97. All calculated values for the 4-CNN model are summarized in Table 3. The ROC curves of the 4-CNN model can be seen in Figure 1. The false positive rate w close to zero, while the true positive rate is between 0.9 and 1.

Table 3. The evaluation metrics for each diagnosis class.

	Accuracy	Precision	Specificity	Sensitivity
AKIEC (class 0)	95%	98%	99%	98%
BCC (class 1)	95%	94%	99%	99%
BKL (class 2)	92%	79%	97%	96%
DF (class 3)	93%	85%	99%	99%
MEL(class 4)	93%	86%	98%	90%
NV (class 5)	94%	97%	97%	91%
VASC (class 6)	93%	98%	99%	98%
Average	93.6%	91%	98.3%	95.9%

10.12865/CHSJ.46.02.06

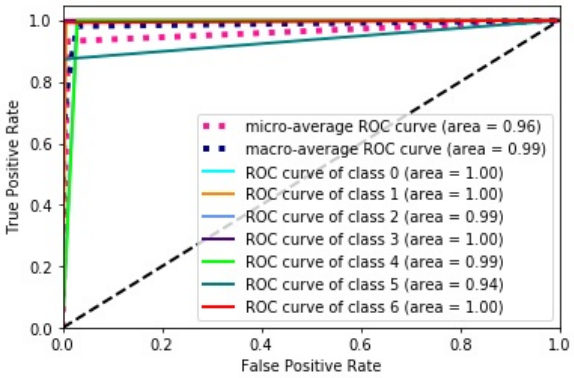

Figure 1. The ROC curves computed for our 4-CNN model.

The 4-CNN precision-recall curves can be observed in Figure 2. The high area under the curve represents very high recall and very high precision. The high precision represents the low false positive rate, and high recall means a low false negative rate.

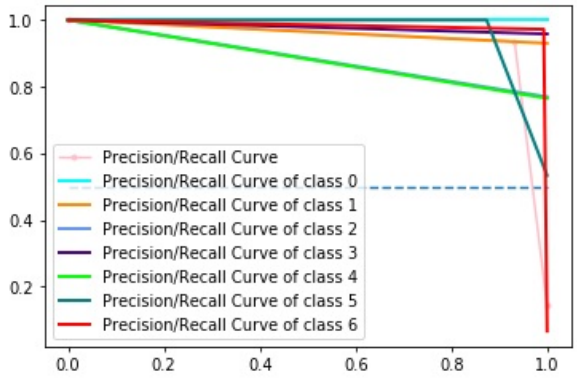

Figure 2.The PR curves computed for our 4-CNN model.

Discussion

In recent years, deep learning methods have proven to be excellent tools in assessing dermatoscopic skin lesions that go beyond classical machine learning methods [5].

Some studies for the classification of skin dermatological lesions were reported in [16-19].

In [16] a study is developed for efficient skin tumour classifier using DCNN (deep convolutional neural network) trained on a relatively small dataset of 4867 clinical images obtained from 1842 patients divided in 14 diagnosis.

They achieved 96.3% sensitivity (correctly classified malignant as malignant) and 89.5% specificity (correctly classified benign as benign).

In [17], the Google Inceptionv3 CNN was trained using a dataset of 129,450 clinical images consisting of 2,032 different diseases. This study did not publish the accuracy obtained but surpassed average-level dermatologists in the sensitivity/specificity.

In [18], the convolutional neural network (Microsoft ResNet-152 model) was trained and tested to classify the clinical images of 12 skin diseases. The model was tested using three images datasets. The following results were obtained: with the Asan dataset, the area under the curve for the diagnosis of basal cell carcinoma, squamous cell carcinoma, intraepithelial carcinoma, and melanoma was approximatively 0.96, 0.83, 0.82, and 0.96, respectively; with the Edinburgh dataset, the area under the curve for the corresponding diseases was 0.90, 0.91, 0.83, and 0.88, respectively.

By comparison with previous studies, our 4-CNN method recorded very good results even with less images to train with an average accuracy of 93.6%, average precision of 91%, average specificity of 98.3%, and average sensitivity of 95.9%.

Our CNN method for skin tumor classification has some limitations:

- the testing dataset was known and there is a risk of overfitting;
- the dataset is imbalanced and has less images for rare tumors;
- the dermatoscopic images were not standardized, having different camera angles, orientations, multiple skin backgrounds, and lighting.

Conclusion

In this study, we focused on finding a way to differentiate the skin lesions, in order to diagnose the dermatoscopic images.

For this purpose, we developed a deep learning algorithm that is superior to the traditional machine learning methods as it does not require an initial description of visual features of medical images.

The proposed 4-CNN model had a good performance in terms of the sensitivity (95.9%), specificity (98.3%), test accuracy (93.6%) and AUC (0.97).

In clinical practice, our deep learning algorithm could assist the dermatologists in the process of diagnosis.

Abbreviations

Akiec Actinic keratoses and intraepithelial carcinoma

Bcc-Basal cell carcinoma

Bkl-Benign lesions of the keratosis

Df-Dermatofibroma

Mel-Melanoma
Nv-Melanocytic nevi
Vasc-Vascular lesions
AI-Artificial intelligence
DL-Deep Learning
CNN-Convolutional neural network
Pre-Precision
Sp-Specificity
Se-Sensitivity
AUC-Area under the curve
ROC-Receiver operating characteristic

Conflict of interests

None to declare.

Acknowledgements

The research leading to these results has received funding from Competitiveness Operational Program 2014-2020 under the project P_37_357 "Improving the research and development capacity for imaging and advanced technology for minimal invasive medical procedures (iMTECH)" grant, Contract No. 65/08.09.2016, SMIS-Code: 103633. This work was also partially supported by the grant POCU 380/6/13/123990, co-financed by the European Social Fund within the Sectorial Operational Program Human Capital 2014-2020.

References

1. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data, 2018, 5:180161.
2. National Cancer Institute, Cancer Stat Facts: Melanoma of the Skin. Available at: https://seer.cancer.gov/statfacts/html/melan.html [Accessed 05.12.2019]
3. Surgeon General's Call to Action to Prevent Skin Cancer. MMWR Morb Mortal Wkly Rep, 2014, 63(30):656.
4. Karagas MR, Weinstock MA, Nelson HH. Keratinocyte carcinomas (basal and squamous cell carcinomas of the skin. In: SchottenfeldD, Fraumeni JF (Eds.): Cancer Epidemiology and Prevention, Oxford University Press 2006, New York, 1230-1250.
5. Fujisawa Y, Inoue S, Nakamura Y. The Possibility of Deep Learning-Based, Computer-Aided Skin Tumor Classifiers. Front Med, 2019, 6:191.
6. Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence estimate of nonmelanoma skin cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. JAMA Dermatol, 2015, 151(10):1081-1086.
7. Deinlein T, Richtig G, Schwab C, Scarfi F, Arzberger E, Wolf I, Hofmann-Wellenhof R, Zalaudek I. The use of dermatoscopy in diagnosis and therapy of nonmelanocytic skin cancer. J Dtsch Dermatol Ges, 2016, 14(2):144-151.
8. Kerr OA, Tidman MJ, Walker JJ, Aldridge RD, Benton EC. The profile of dermatological problems in primary care. Clin ExpDermatol, 2010, 35(4):380-383.
9. Keras. Available at: https://github.com/keras-team/keras. [Accessed 05.07.2020].
10. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 2014, 15(1):1929-1958.
11. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R. Tensor Flow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Available at: https://arxiv.org/pdf/1603.04467.pdf. [Accessed 10.07.2019].
12. Walt SV, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Computing in Science and Engineering, 2011, 13(2):22-30.
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: Machine Learning in Python. J Mach Learn Res, 2011, 12:2825-2830.
14. Brownlee J. Deep Learning with Python, Machine Learning Mastery. Available at: https://machinelearningmastery.com/machine-learning-with-python. [Accessed 02.10.2019].
15. Saito, T, Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One, 2015, 10(3):e0118432.
16. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, Watanabe R, Okiyama N, Ohara K, Fujimoto M. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol, 2019, 180(2):373-381.
17. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542 (7639):115-118.
18. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol, 2018, 138(7):1529-1538.
19. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, Schilling B, Haferkamp S, Schadendorf D Fröhling S, Utikalh J.S, von Kalle C.A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer, 2019, 111:148-154.

Corresponding Author: Lucian Gheorghe Gruionu, Faculty of Mechanics, University of Craiova, Craiova, Romania, e-mail: lgruionu@gmail.com