Baltic J. Modern Computing, Vol. 10 (2022), No. 2, 241-250
https://doi.org/10.22364/bjmc.2022.10.2.10

Classification of Skin Lesion Images with Deep Learning Approaches

Buket BAYRAM[1], Bahadır KULAVUZ[2], Berkay ERTUĞRUL[2],
Bulent BAYRAM[2], Tolga BAKIRMAN[2], Tuna ÇAKAR[3],
Metehan DOĞAN[4]

1 Dermatology Clinic (M.D.), Istanbul, Turkey
2 Yildiz Technical University, Faculty of Civil Engineering,
Department of Geomatics Engineering, Istanbul, Turkey;
3 Graduate School of Big Data Analytics, MEF University, Istanbul, Turkey
4 Software Engineer in BeGeo Software Tech. Inc. Co. Sakarya, Turkey

drbuketbayram@gmail.com, bahadırkulavuz@gmail.com,
berkayerturul80@gmail.com, bayram/bakirman@yildiz.edu.tr,
tuna.cakar@tunasc.com, metehandogn@gmail.com

ORCID 0000-0001-9961-2151, 0000-0002-7269-5708, 0000-0002-4248-116X,
0000-0001-7828-9666, 0000-0001-8594-7399, 0000-0002-3711-8315

Abstract. Skin cancer is one of the most dangerous cancer types in the world. Like any other cancer type, early detection is the key factor for the patient's recovery. Integration of artificial intelligence with medical image processing can aid to decrease misdiagnosis. The purpose of the article is to show that deep learning-based image classification can aid doctors in the healthcare field for better diagnosis of skin lesions. VGG16 and ResNet50 architectures were chosen to examine the effect of CNN networks on the classification of skin cancer types. For the implementation of these networks, the ISIC 2019 Challenge has been chosen due to the richness of data. As a result of the experiments, confusion matrices were obtained and it was observed that ResNet50 architecture achieved 91.23% accuracy and VGG16 architecture 83.89% accuracy. The study shows that deep learning methods can be sufficiently exploited for skin lesion image classification.

Keywords: Deep Learning, Image classification, ISIC 2019, ResNet50, VGG16

1. Introduction

Cancer is defined as a disease consisting of uncontrolled proliferation of foreign cells in organs (Cooper, 2019). Cancer is the highest cause of death in the world and skin cancer is one of the most common types of cancer in the world (WHO, 2022). Skin cancer is one of the three most dangerous types of cancer caused by damaged DNA that can cause death (Ali et al., 2021). Early detection of skin cancer can highly increase the curing rate (Codella et al., 2017). Since the time-consuming visual examination of the lesions is dependent on the dermatologist's prior experience, this can cause the disease to be misdiagnosed (Kazakevičiute-Januškevičiene et al., 2015).

Computer-aided diagnostic systems help clinicians improve the accuracy of their diagnoses by providing a second perspective (Yanese and Triantaphyllou, 2019). In this context, utilization of image processing and deep learning algorithms can increase dermatologist performance and minimize the diagnostic time in the detection of skin cancer (Hosny et al. 2019).

The diagnosis and detection process in medicine with convolutional neural networks has increased in use in recent years. Gessert et al. (2020) have used ensembles of multi-resolution EfficientNet to classify ISIC 2019 datasets that have achieved results with 74.2% sensitivity. Mahbod et al. (2020) have used EfficientNet architectures and ISIC 2017 dataset to explain the impact of segmentation on classification. Three different segmentation models were used and the results of manual segmentation achieved the highest classification accuracy with an area under the receiver operating characteristic curve (AUC) score of 93%. Harangi et al. (2020) created a supported deep learning framework using the GoogleNeT Inception-v3 architecture and performed a seven-class classification with an accuracy score of approximately 90% for each class. Sekhar et al. (2021) have used raw dermoscopic images as an input to the CNN and features of segmented dermoscopic images as additional information. The proposed method gives a classification accuracy of 98.13% for the identification of Melanoma. Maron et al. (2021) have implemented VGG16_BN, ResNet50, DenseNet121 and AlexNet architectures to test the robustness of convolutional neural networks in skin cancer using 3 different datasets (Skin Archive Munich (SAM), SAM-corrupted (SAM-C) and SAM-perturbed (SAM-P)). Calderon et al. (2021) performed classification on the HAM10000 dataset to compare the state-of-the-art architectures with the bilinear approach created in the VGG16 and ResNet50 architectures. It was seen that the new approach achieved higher accuracy than other methods, with an F1 score of 0.9321. Hasan et al. (2022) have used a hybrid convolutional neural network (DermoExpert) to classify ISIC 2016, ISIC 2017 and ISIC 2018 datasets that have achieved the area under the receiver operating characteristic curve (AUC) of 0.96, 0.95, and 0.97, respectively. Indraswari et al. (2022) have used MobileNetV2 network to classify melanoma datasets and achieved an accuracy of over 85%.

With these algorithms, results as successful as an expert can be obtained and human error can be eliminated (Kassam, 2016). Therefore, we performed classification using images from the largest published dermoscopic open datasets - the International Collaboration on Skin Imaging (ISIC Archive, 2019) dataset. Due to their success in the literature, we have chosen VGG16 and ResNet50 architectures for dermoscopic image classification since the depth of models is quite different.

In terms of deep learning-based dermoscopic image classification, this study mainly aims to answer the following questions:

- Can architectures with different depths achieve similar accuracy results on the same dataset?
- How does the balance of the dataset affect accuracy?
- Does the accuracy of the architectures increase as the dataset grows?

2. Materials and Methods

2.1. ISIC 2019

The dataset used in the research was obtained from the ISIC 2019 (ISIC Archive, 2019) challenge. There are 9 classes in the dataset and a total of 25331 image data. Classes in the dataset are specified as Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Nevus (NV), Squamous Cell Carcinoma (SCC), Vascular Lesion (VASC) and None of The Others (UNK). A part of the ISIC dataset was obtained by cropping the lesion areas of the images in the HAM10000 dataset at 600x450 sizes, and histogram corrections were applied to some images (Tschand et al., 2018). There are also different datasets consisting of skin lesion images. BCN_20000 and MSK datasets are a few of them. The images in the BCN_20000 are referred to as difficult dataset since the datasets consist of lesions that occur in rare regions, and the size of the images is 1024x1024 (Combalia et al., 2019). The images in the MSK dataset do not have a fixed size. On the other hand, it contains additional information such as the patient's age group, gender, and the region of the lesion. However, since these data are missing in some images, the dataset cannot be used in its full form (Gessert et al., 2020).

While determining the classes to be used in the research, homogeneity was taken into consideration in the data distribution. Among these classes, BCC, MEL and NV classes with the highest number of images were selected for this project. In the dataset, BCC class has the minimum number of images, which is 3323. For this reason, in the first stage of the research, the number of images for 3 classes was equalized to 3323 for a balanced dataset. The dataset created with these images was used in the processes conducted with Dataset 1. For the second part of the study, the number of images was increased using augmentation to 10911 which is the highest number of images for a class (NV). The second dataset created was used in the Dataset 2 phase. The Dataset 2 were generated using augmentation techniques such as horizontal flip, vertical flip and random brightness contrast augmentation.

The number of original images, Dataset 1 images and Dataset 2 images belonging to the classes in the dataset are shown in Table 1. Examples from the classes in the dataset are shown in Figure 1.

Table 1. Number of images

Class	Original Image Count	Dataset 1	Dataset 2
BCC	3323	3323	10911
NV	10911	3323	10911
MEL	3970	3323	10911
Total	18204	9969	32733

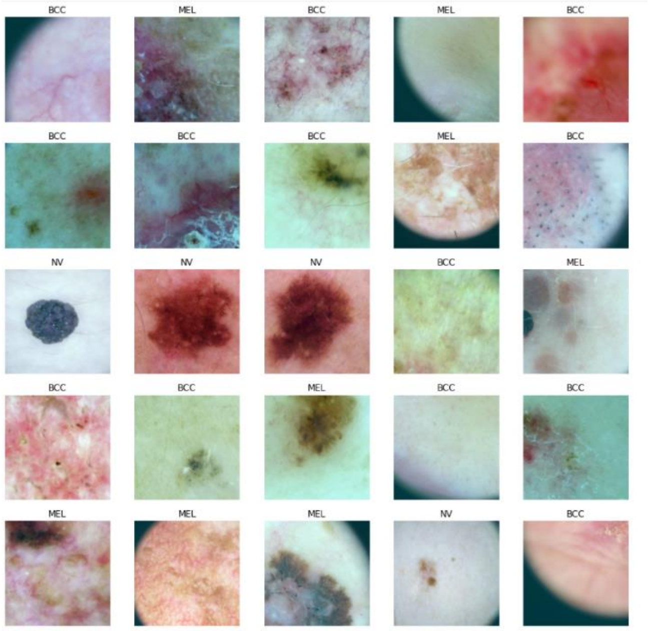

Figure 1. Original Examples from ISIC 2019 Dataset (ISIC Archive, 2019)

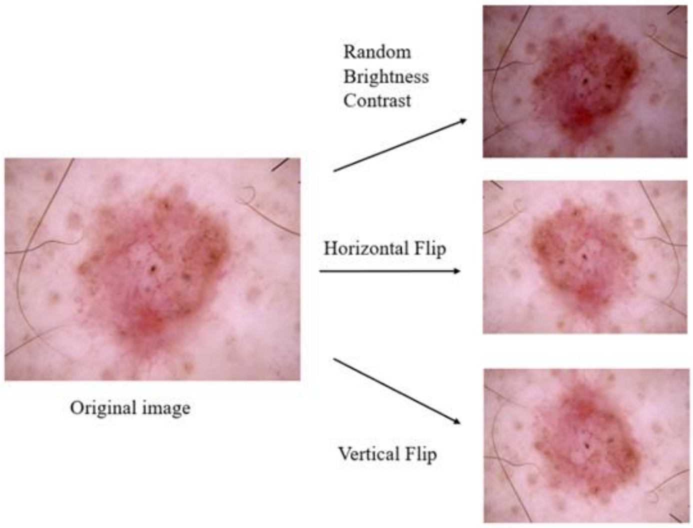

Figure 2. Random Brightness Contrast, Horizontal and Vertical Augmentation Examples

2.2. Architectures

In this study, ResNet and VGG architectures were used to classify skin lesion images. ResNet (He et al., 2016) architecture made its name by winning the ImageNet Classification challenge held in 2015. ResNet50 architecture consists of 48 convolutions, 1 maximum pooling and 1 average pooling layer. The input image size of the architecture is 224x224 and the first layer is 7x7. It is the convolution layer with kernel size, and then there is the max-pooling layer with a 3x3 kernel size. The most important feature of the ResNet architecture is that it uses the residual learning method in the learning process for which it exploits Residual blocks (Figure 2). Residual blocks consist of 3 convolution layers and these layers have kernel sizes of 1x1, 3x3 and 1x1, respectively. Finally, after the residual blocks, there is an average pooling and a fully connected layer with 1000 neurons (Calderon et al., 2021).

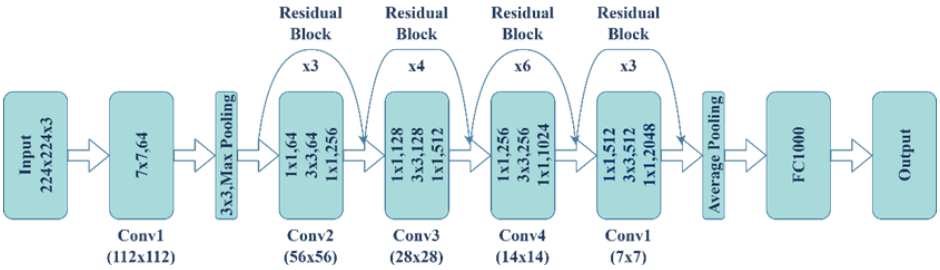

Figure 3. ResNet50 Architecture

The VGG architecture was first introduced by Simonyan and Zisserman (2015) in the ImageNet Classification Challenge. The VGG16 architecture consists of 16 layers, including 13 convolutional layers and 3 fully connected layers. The convolution layers form 5 convolution blocks and the input size is 224x224 pixels (Figure 3). There are convolution and maximum pooling layers within the convolution blocks. The convolution layer kernel size is 3x3, the maximum pooling kernel size is 2x2, and the fully connected layer kernel size is 1x1. After the convolution blocks, there are 3 fully connected layers with 4096, 4096 and 1000 neurons. (Göçeri, 2019). SoftMax activation is located on the final layer as can be seen in Figure 3.

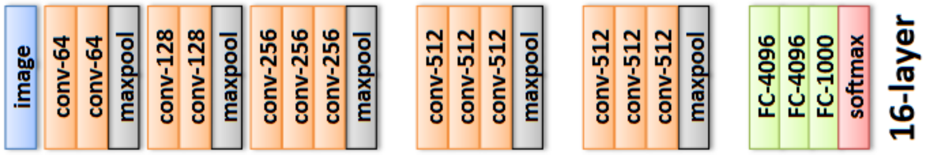

Figure 4. VGG16 Architecture (Simonyan and Zisserman, 2015)

It is aimed to perform the mentioned deep learning approaches for lesion classification from dermoscopic images which can be captured in various angles and lightening conditions.

Information about the equipment on which the trainings are carried out is shown in Table 2. The hyperparameter information of the architectures is shown in Table 3 which are determined empirically considering hardware limitations.

Table 2. Hardware specifications

Hardware specifications	
Operating System	Ubuntu 20.04.3 LTS
Processor	Intel® Xeon(R) Silver 4214 CPU @ 2.20GHz × 48
RAM	126 GB
Graphics Card	NVIDIA GeForce RTX 2080 Ti

Table 3. Information of hyperparameters

Hyperparameter	ResNet50	VGG16
Library	PyTorch	PyTorch
Batch Size	8	16
Activation Function	Softmax	ReLU
Optimizer	Adam	Adam
Learning Rate	5e-04	1e-04
Number of Epoch	100	100

3. Result and Discussion

The datasets have been split as 70%, 20% and 10% for train, validation and test sets, respectively. In Dataset-1, the number of images is 6978, 1993 and 997 for train, validation and test sets, respectively. In Dataset-2, the number of images is 22913, 6547 and 3273 for train, validation and test sets, respectively.

Overall accuracy test results of VGG-16 and ResNet-50 architecture are shown in Table 4. The test datasets of Dataset 1 and Dataset 2 were also cross-compared with the trainings conducted with both datasets. For example, both networks are trained with Dataset-1 and Dataset-2, and tested with the test set of Dataset-1 and Dataset 2. The results show that increasing the number of images in Dataset 2 does not affect the performance of VGG16 as the overall accuracy increased around only 1% for the test set of Dataset-2. On the contrary, ResNet50 performed significantly better when it is trained with Dataset-2 since the overall accuracy is increased by almost 10% for the test set of Dataset-2. For both networks, increasing number of images in the training (Dataset-2) does not seem to perform well on the small test set of images (Dataset-1), since the overall accuracy of both networks is decreased by approximately 14% and 13% for VGG16 and ResNet50, respectively.

The confusion matrix (Table 5) shows the relationships between the test classes as a result of the prediction of the data whose real classes are known. Additionally, Table 6 shows precision, recall and F1 values for each class in all experiments. The observations on both tables show that the class that was confused the most and reduced the accuracy as a result of tests was the MEL class as it has the lowest F1 value. It was observed that the BEL and NV classes were better distinguished. In terms of F1 values, ResNet50 has outperformed VGG16 for all classes in all experiments. In line with the overall accuracy, the best results seem to be obtained with ResNet50 trained with Dataset-2. Some TP examples are shown in Figure 4.

Table 4. VGG16 and ResNet50 architecture overall accuracy test results. The underlined values indicated the best results for each architecture and dataset

		Test	
Train	**VGG16 (Accuracy %)**	Dataset-1	Dataset-2
	Dataset-1	83.89	76.37
	Dataset-2	69.66	77.38
	ResNet50 (Accuracy %)	Dataset-1	Dataset-2
	Dataset-1	83.35	81.45
	Dataset-2	70.52	91.23

Table 5. Confusion Matrix for all four cases for each architecture

Train/Test: Dataset-1/Dataset-1							
VGG16	BCC	MEL	NV	**ResNet50**	BCC	MEL	NV
BCC	592	56	17	BCC	570	80	15
MEL	127	474	64	MEL	92	521	47
NV	24	42	599	NV	20	57	588
Train/Test: Dataset-2/Dataset-1							
VGG16	BCC	MEL	NV	**ResNet50**	BCC	MEL	NV
BCC	571	37	57	BCC	649	2	14
MEL	101	316	248	MEL	66	312	287
NV	18	23	624	NV	1	1	663
Train/Test: Dataset-1/Dataset-2							
VGG16	BCC	MEL	NV	**ResNet50**	BCC	MEL	NV
BCC	1864	273	46	BCC	1867	243	73
MEL	463	1544	176	MEL	386	1601	196
NV	342	659	1182	NV	312	717	1154
Train/Test: Dataset-2/Dataset-2							
VGG16	BCC	MEL	NV	**ResNet50**	BCC	MEL	NV
BCC	1967	121	95	BCC	2107	63	13
MEL	484	1231	468	MEL	140	1772	271
NV	174	142	1862	NV	62	26	2095

Considering the confusion matrices (Table 5) and incorrectly predicted images (Figure 5), MEL class should be more investigated. Even though we have created a balanced dataset with an equal number of images for each class, it does not look sufficient for MEL classification.

Table 6. Precision, Recall and F1 values of each experiment for each class. The best F1 values for each class are underlined.

Train/Test: Dataset-1/Dataset-1							
VGG16	Precision	Recall	F1	**ResNet50**	Precision	Recall	F1
BCC	0.8902	0.7968	0.8409	BCC	0.8571	0.8358	0.8463
MEL	0.7128	0.8287	0.7664	MEL	0.7894	0.7918	0.7906
NV	0.9008	0.8809	0.8907	NV	0.8842	0.9046	0.8943
Train/Test: Dataset-2/Dataset-1							
VGG16	Precision	Recall	F1	**ResNet50**	Precision	Recall	F1
BCC	0.8587	0.8275	0.8428	BCC	0.9759	0.9064	0.9399
MEL	0.4752	0.8404	0.6071	MEL	0.4692	0.9905	0.6367
NV	0.9383	0.6717	0.7829	NV	0.9970	0.6877	0.8140
Train/Test: Dataset-1/Dataset-2							
VGG16	Precision	Recall	F1	**ResNet50**	Precision	Recall	F1
BCC	0.8539	0.6984	0.7683	BCC	0.8551	0.7279	0.7864
MEL	0.7072	0.6236	0.6628	MEL	0.7334	0.6251	0.6750
NV	0.5415	0.8419	0.6591	NV	0.5286	0.8110	0.6400
Train/Test: Dataset-2/Dataset-2							
VGG16	Precision	Recall	F1	**ResNet50**	Precision	Recall	F1
BCC	0.9010	0.7493	0.8182	BCC	0.9652	0.9125	0.9381
MEL	0.5639	0.8240	0.6696	MEL	0.8117	0.9522	0.8764
NV	0.8549	0.7678	0.8090	NV	0.9597	0.8806	0.9185

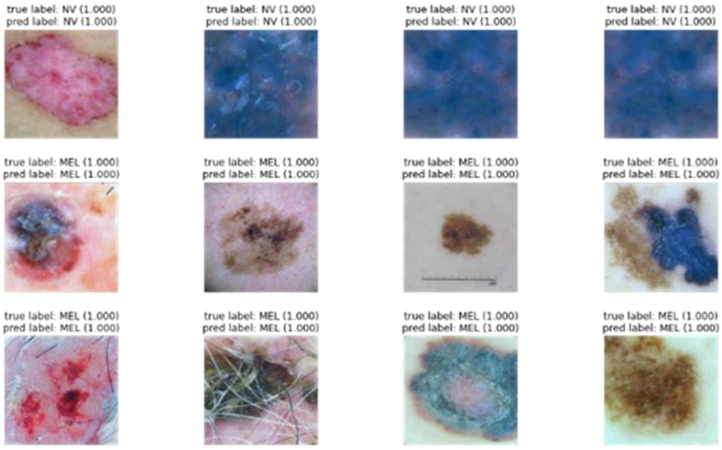

Figure 5. Examples of the correct prediction images

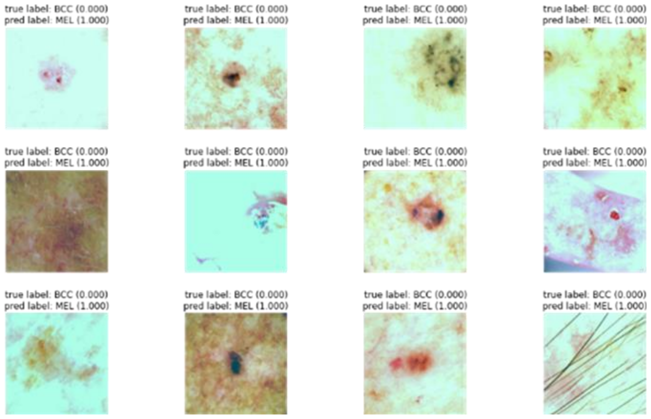

Figure 6. Examples of incorrectly predicted images of VGG architecture

4. Conclusion

As a result of the study, it was observed that the accuracy of the ResNet50 architecture increased in parallel with the number of image data. However, when the accuracies obtained with the VGG-16 architecture were examined, it was observed that the accuracy is decreased as a result of the training with more data. For this reason, it has been determined that VGG-16 does not provide the desired performance in our data set.

The performance of ResNet50 can be improved even further in terms of MEL classification by increasing the number of images. For future studies, we aim to perform recently popular visual information transformers and semantic segmentation of skin lesions to extract morphology and boundary. Thus, a semantic segmentation-based computer-aided diagnosis approach will be developed to give physicians a second opinion for improvement of their diagnosis.

References

Ali S., Miah S., Haque J., Rahman M., Islam K. (2021). An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models, *Machine Learning with Applications*, **5**, 100036.

Calderon C., Sanchez K., Castillo S., Arguello H. (2021). BILSK: A bilinear convolutional neural network approach for skin lesion classification, Computer Methods and Programs in Biomedicine Update 1

Codella, N., Nguyen, Q. B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., Smith, J. R. (2018). Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. *IBM Journal of Research and Development* **61**(5)**,** 1-15

Combalia, M., Codella, N. C. F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A. C., Puig, S., Malvehy J. (2019). BCN20000: Dermoscopic Lesions in the Wild, *arxiv: 1908.02288*

Cooper, G. M. (2019) *The Cell: A Molecular Approach*, Sinauer Associates, Oxford University Press

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**

Goçeri E., (2019), Analysis of Deep Networks With Residual Blocks and Different Activation Functions: Classification of Skin Diseases. *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA),* IEEE, pp. 1–6.

Harangi B., Baran A., Hajdu A., (2020), Assisted deep learning framework for multi-class skin lesion classification considering a binary classification support, *Biomedical Signal Processing and Control* **62**

Hasan, K. ,Elahi, T. E., Alam, A., Jawar, T., Marti R. (2022), DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation, *Informatics in Medicine Unlocked*, 28, 100819.

He K., Zhang X., Ren S., Sun J. (2016), Deep Residual Learning for image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770–778.

Hosny, K. M., Kassem, M. A., Foaud, M. M. (2019) Skin Cancer Classification using Deep Learning and Transfer Learning, *9th Cairo International Biomedical Engineering Conference (CIBEC),* 20-22 Dec. 2018, Cairo, Egypt

Indraswari R., Rokhana R., Herulambang W. (2022), Melonoma image classification based on MobileNetV2 network, *Sixth Information Systems International Conference (ISICO 2021)*

ISIC Archive. International skin imaging collaboration: Melanoma project website [Online] .https://isic-archive.com, 2019.

Kassam, A. (2016), Segmentation Of Skin Cancer By Using Image Processing Techniques, Master Thesis, Yıldız Technical University Department Of Computer Engineering, İstanbul

Kazakevičiute-Januskevičiene, G., Ušinskas, A., Januškevičius, E., Ušinskiene, J. (2015) Region-based Annotations for the Medical Images, *Baltic Journal of Modern Computing* **3**(4), 248-267

Sekhar, K. S. R., Tummala, R. B., Goriparthi, P., Kotra, V. (2021), Dermoscopic image classification using CNN with Handcrafted features, *Journal of King Saud University – Science* **33**

Mahbod A., Tschandl P., Langs G., Ecker R., Ellinger I. (2020). The effects of skin lesion segmentation on the performance of dermatoscopic image classification, *Computer Methods and Programs in Biomedicine* **197**

Maron, R.C., Schlager, J.G., Hanggenmüller, S., Kalle, C. von, Utikal, J.S., Meier, F., Gellrich, F.F. et al. (2021), A benchmark for neural network robustness in skin cancer classification, *European Journal of Cancer* **155,** 191-199.

Simonyan K., Zisserman A., (2015), Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015

Tschandl P., Rosendahl C., Kittler H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, (eng), *Scient. Data* **5,** 180161.

Yanese, J., Triantaphyllou, E. (2019) A systematic survey of computer-aided diagnosis in medicine: Past and present developments, *Expert Systems with Applications* **138**, 112821.

WHO (2022) World Health Organization – Cancer, https://www.who.int/news-room/fact-sheets/detail/cancer *Access Date: 10 Feb 2022*

Received May 31, 2022, accepted June 17, 2022