Deep Melanoma classification with K-Fold Cross-Validation for Process optimization

Yali Nie
Dept. Electronics Design
Mid Sweden University
Sundsvall, Sweden
yali.nie@miun.se

Laura De Santis
Dept. Industrial Engineering
University of Salerno
Salerno, Italy
ldesantis@unisa.it

Marco Carratù
Dept. Industrial Engineering
University of Salerno
Salerno, Italy
mcarratu@unisa.it

Mattias O'Nils
STC Research Center
Mid Sweden University
Sundsvall, Sweden
mattias.onils@miun.se

Paolo Sommella
Dept. of Industrial Engineering
University of Salerno
Salerno, Italy
psommella@unisa.it

Jan Lundgren
STC Research Center
Mid Sweden University
Sundsvall, Sweden
Jan.Lundgren@miun.se

***Abstract*—Deep convolution neural networks (DCNNs) enable effective methods to predict the melanoma classes otherwise found with ultrasonic extraction. However, gathering large datasets in local hospitals in Sweden can take years. Small datasets will result in models with poor accuracy and insufficient generalization ability, which has a great impact on the result. This paper proposes to use a K-Fold cross validation approach based on a DCNN algorithm working on a small sample dataset. The performance of the model is verified via a Vgg16 extracting the features. The experimental results reveal that the model built by the approach proposed in this paper can effectively achieve a better prediction and enhance the accuracy of the model, which proves that K-Fold can achieve better performance on a small skin cancer dataset.**

Keywords—DCNNs, melanoma, K-Fold, classification

I. Introduction

Skin cancer has been increasing throughout the world recently. The World Health Organization estimates that there are about 3 million skin cancer cases found globally every year [1]. There 54% increase in the number of new invasive melanoma cases diagnosed each year over the past decade (2009-2019) [25]. Melanoma is serious type of skin cancer, since it is the main cause for a majority of the mortalities caused by skin cancer. The Australian Government's website [2] shows that melanoma skin cancer was taken as the fourth most common cancer diagnosis in 2019. There was a total of 15,229 new cases of melanoma skin cancer in Australia that year. Of the affected, 8,899 were male patients and 6,330 were female. There were 1,725 deaths from melanoma skin cancer. According to the American Cancer Society [3], the annual incidence of melanoma skin cancer increases by approximately 3% per year. Though melanoma is one of the most serious type of skin cancer in the world, the chance of recovering from this cancer is high if it is diagnosed at an early stage.

Over the past decade, most image classification techniques utilize classical machine learning methods to extract features relying on hand-crafted features. Aitken et al. classified pigmented skin lesions, by including their color, size, shape, and distinctness of boundary [13]. Based on the criteria asymmetry (A), border (B), color (C), and differential structure (D), Nachbar et al. developed the "ABCD" [11] rule to make classification of pigmented skin lesions. Yuan et al. employed a support vector machine (SVM) based on texture information [12] to achieve early melanoma detection. The 7-point checklist [15] is a typical method for diagnosis of melanoma skin cancer [26, 27, 28]. Here, each image has its individual scores which will make a simple addition and a minimum total score of 3 out of 7 is required for the diagnosis of malignant cell changes, whereas a total score of less than 3 is indicative of benign cell changes.

Recently, DCNNs have become a popular approach to handle the different medical image analysis tasks [6, 7, 8, 9, 10, 22, 23, 24]. They have the ability to learn more complex feature hierarchies from raw data. In 2012, image classification made a big progress since ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) [14]. 2012 marked the first year where a DCNN was used to make an achievement with a top 5 test error rate of 15.4%. However, the next best entry achieved an error rate of 26.2%, that was a remarkable improvement that amazed the computer vision community.

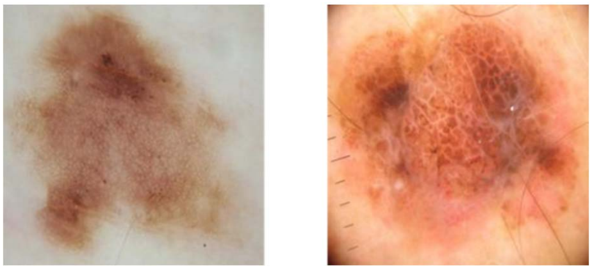


Fig. 1. Samples of benign and malignant melanoma from our dataset.

In this work, the researchers in collaboration with the medical community built their own dataset, since many publicly available skin lesion datasets are imbalanced, and sometimes acquired in an unprofessional manner. Fig. 1 shows two samples of cell changes in the skin with benign and malignant melanoma from the acquired dataset. A DCNN was trained with 5-Fold cross validation techniques to classify images of suspect dermoscopy lesions as benign or malignant. Dermoscopy images have played a extraordinary role in the medical area for diagnosis of skin cancer by increasing the survival rate of patients. It is an efficient skin imaging technique that allows a magnified visualization of skin surface [4]. This study aims to develop a DCNN that can automatically classify the melanoma skin cancer on a small dermoscopy dataset.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©2020 IEEE

978-1-7281-5386-5/20/$31.00 ©2020 IEEE

The layout of the paper continues with the connected work in Section II. And in Section III, the research methodology is presented. The experiment and result related to data of melanoma skin cancer, as well as the training process and the analysis of the experiment results are demonstrated in Section IV. The last section presents the conclusions and future plans within this field of research.

II. Related Work

With the increasing computational power of GPUs in mind, the melanoma classification model was built based on a deep learning algorithm and K-Fold cross validation to make a classification of melanoma skin cancer. Since the dataset is small, this paper use cross-validation to prevent the network from over-fitting.

A. *Deep learning*

Deep learning is one technique from machine learning that utilizes many layers of nonlinear information processing to perform feature extraction, pattern recognition, and classification [19]. Deep neural networks have more than one hidden layer. One simple neural network is shown in Fig. 2. It is a four-layer network that combines with two hidden layers, one input layer and one output layer. Although the design of the input and output layers of a neural network is usually simple, the design of hidden layers can be an art. In deep learning, computer learning classifies images directly [14], text [20], or a sound [21]. DCNNs is one deep learning algorithm which is very efficient for solving general and highly variable tasks with large datasets. It can extract the regional characteristics of the original image, based on the local feature extraction. Generally, by adding more layers, the learning model can get high precision through Supervised Learning. Just as a computer is trained to use large datasets and then change the pixel value of an image to an internal representation or vector feature, where classifiers can detect or classify patterns in the input [18]. Simply said, the CNNs are composed of feature extraction layers and the classification layer. However, they can efficiently extract the regional features of the original image based on the local information. CNN is thus a method for transforming the original image, layer by layer, from the image pixel value into the class scoring value for classification, where each layer has a hyper parameter and some do not have parameters (weight and bias on neurons) [7].

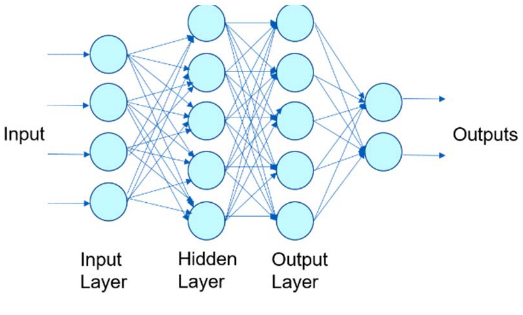


Fig. 2. An architecture of a simple neural network.

B. *K-Fold cross validation*

Cross-validation is one statistical method which can be used to estimate machine learning models' skills. It is frequently used in involved machine learning to make a comparison and select models for a given predictable modeling problem, since it is easy to learn, easy to carry out, and results in skills estimate that it generally has a lower bias than other methods as the mean estimate of any parameter is less biased than a one-shot estimate [17]. According to paper [16], the whole process of the K-Fold cross validation algorithm is shown in the following four steps:

1. Split the entire training set S into k subsets evenly. Assuming that the whole number of training examples in S sets m, thus each subset has m/k training examples. The corresponding subset is denoted as $\{S_1, S_2, \dots, S_k\}$.

2. Take one from the model set as M_i, Then select k-1 in the training subset $\{S_1, S_2, S_{j-1}, S_{j+1,} \dots, S_k\}$. (That is, only one is left as S_j). Do training with this k-1 subset M_i, and then get the hypothetical function h_{ij}. Take the remaining one S_j as test and get experience errors $\hat{\varepsilon}_{S_j}(h_{ij})$.

3. Since we leave one S_j at a time (j from 1 to k), we get k empirical errors. Thereby, for an M_i, its empirical error is the average of these k empirical errors. The performance indicator reported by the K-Fold cross-validation is then the average of the values calculated in the loop.

4. Pick the one with the lowest average empirical error rate M_i, Then use all S to do another training to get the final h_i.

Test the performance of the model through the above steps 1, 2 and 3, and take the average value as the performance index of a model. Get the fusion of all trained K-Folds by calculating the overall performance of all models.

This algorithm can be computationally costly, but does not squander too much data. That is a major advantage to solve some problems such as inverse inference where the number of samples is in small size.

III. Research Methodology

A. *Data Collection*

We focus on dermoscopic images which allow the clinician to perform higher diagnostic accuracy for the melanoma classification compared with traditional eye examination [5]. The dataset is obtained from the University of Salerno in collaboration with the University of Naples, where the dataset has been labeled accurately by experts. Though there are only around 1000 images, all of them are very high-quality annotations. In order to have the balanced melanoma classification, in other words, to guarantee that the proportion of positive and negative examples are the same in the training set, the images that constitutes benign and malignant are set half and half. The total number is 760 for training. There are another 191 images for testing. Each image is named in the format as the diagnosed name by expert with a numeric serial number. The known names will be used as background truth during training.

B. *Training Data Process*

The training process starts by reading the dataset with benign and malignant. Then all the images with different resolutions are resized into 224 x 224 as a standard VGG16 size. The training dataset is split into K subsets, which is 152 per subset for cross-validation, instantiating K identical models. Here each model is trained on K-1 partitions and evaluated on the remaining partitions. The process has a single parameter called K which represents the number of groups into which a given data sample is split. In the model, K is 5 and the processing is shown in Fig. 3. The final evaluation performance of K times is averaged as the overall performance of the algorithm. To improve the results with this small dataset, it is necessary to apply cross validation.

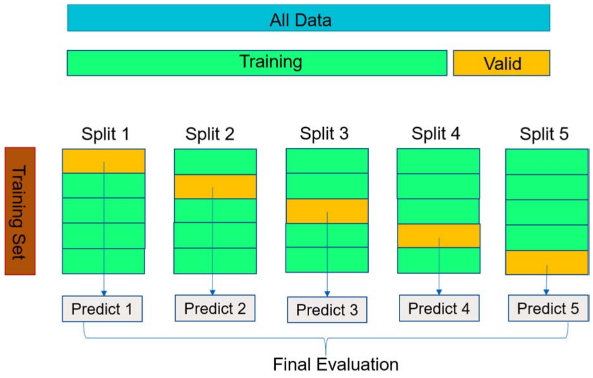


Fig. 3. The process of 5-Fold cross validation.

The system initializes Vgg16 architecture [9] and starts to train the network with the epoch set manually. This CNN framework includes multiple convolution layers, multiple filters, max pooling layers, dropout layers, softmax layers, and an activation function. A non-liner activation function (ReLU) is applied to effectively solve the problem of gradient disappearance. In the end, the training will give a probability value for the two classes, where the predicted class will be compared with the labeled data. The weights will change depending on the ground truth. Finally, the training weights are stored in the form of a model file.

IV. Experiment and Result

This work on the classification of melanoma cancer images is done in two phases. For the first stage, the dataset is trained to produce models and save the best model. In the second stage, the system is produced which initializes the model from the results of the training process. It tests the model with test images and displays the prediction of the classes along with their probabilities.

A. *Experiments*

There are parameters running constantly throughout the procedure in the training, such as batch size and learning rate. In this paper, the batch size is 32 and the learning rate is 0.00001 respectively. The batch size is determined by the memory capability of the device used in the training process. Our device is a Tesla P100. Our environment configuration is installed with a Keras deep learning framework which is an advanced neural network library written in Python and able to run on TensorFlow. An RMSprop [29] optimizer does the optimization of the loss algorithm, which has been growing in popularity in recent years. We use accuracy and loss to measure the algorithm's performance. Accuracy is the measure of how accurate your model's prediction is compared to the true data, however, a loss is not a percentage. It is a sum of the errors made for each example in training or validation sets.

The first layer filters feature map learned from the dataset can be seen in Fig. 5. It shows 32 learned filters feature map and each presents a learned filter with a 3x3 kernel size. The output feature maps are used as the input of the next layer. One of the original input images is shown in Fig. 4.

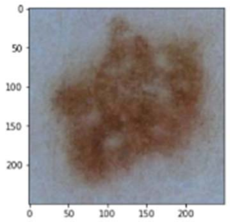


Fig. 4. One example image from input.

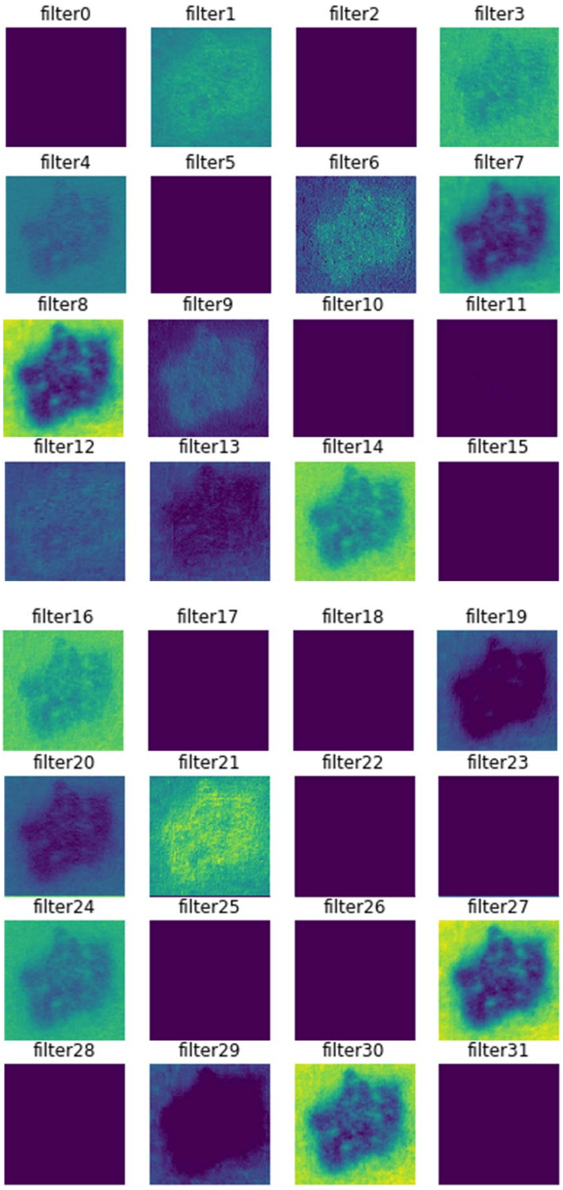
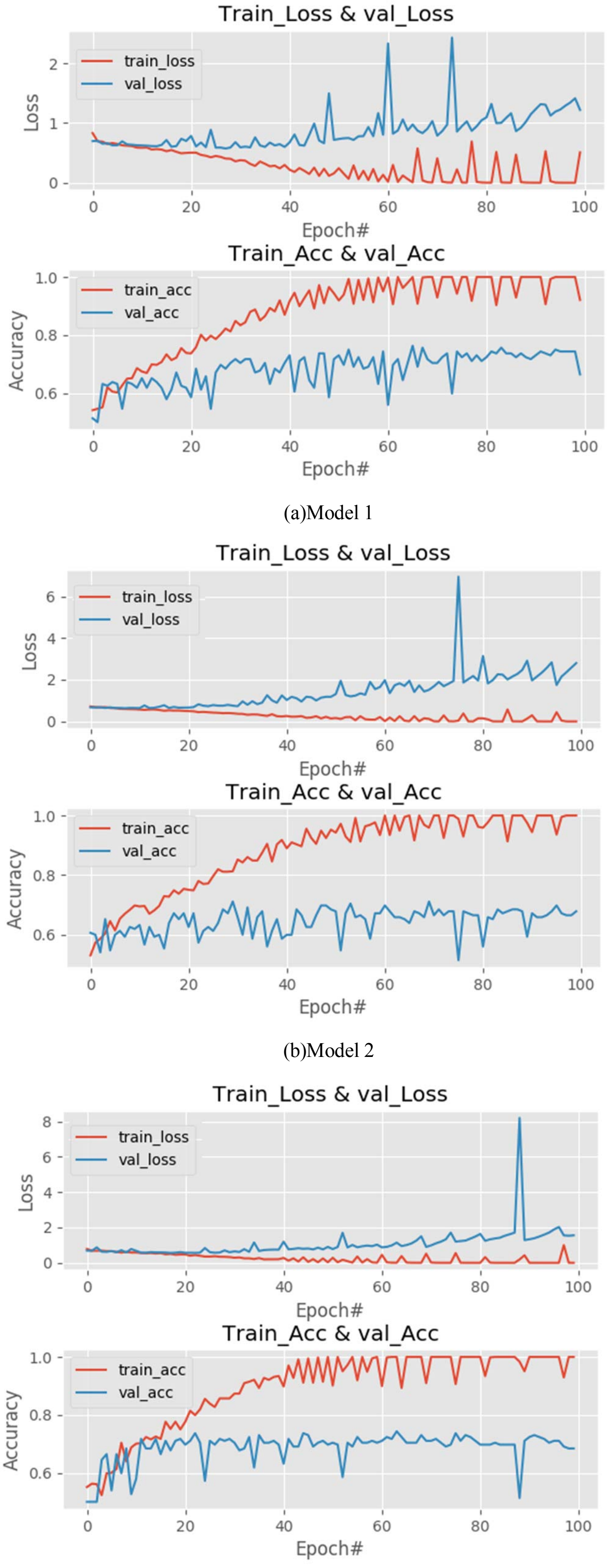


Fig. 5. Visualization of the learned filter from the first convolution layer.

B. *Results*

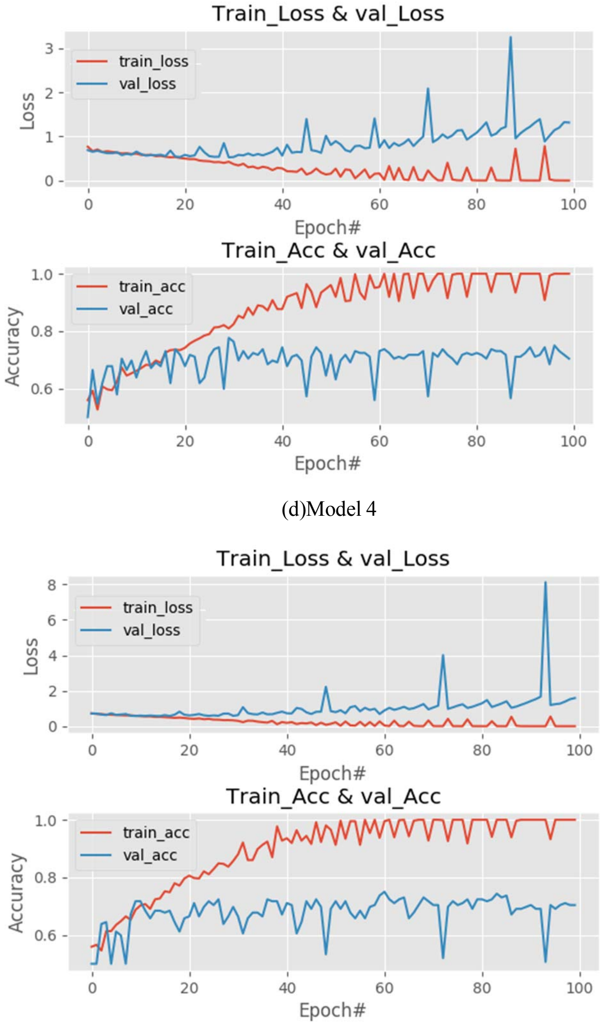
In this paper, the dataset is split into 5 equally sized folds, 5 models are trained and each fold is given an opportunity to be used as the holdout set where the model is trained on all remaining folds. The experiment training process was carried out with 100 epochs on 608 train data and 152 test data which consist of benign and malignant samples in the same number for image balance. The training results of the 5 models are shown in Fig. 6. Line plots are also created showing the learning curves of the model and test set of the training curve at each training period. It can be seen that from epoch 0 to 60 the training accuracy has increased with the final results almost reaching 1 while the validation stays around 0.7. The training loss starts to have a big fluctuation after epoch 60.



(a)Model 1

(b)Model 2

(c)Model 3



(d)Model 4

(e)Model 5

Fig. 6. Each training result of 5-fold with 760 images and 100 epochs. Epoch is a hyperparameter which is total times of an entire dataset passed both forward and backward through the neural network. Accuracy is the measure of how accurate your model's prediction is compared to the true data. Loss is a sum of the errors made for each example in training or validation sets.

The separated model, resulted from the 5-Fold training, was tested for each of its self-validation parts and calculated for the percentage of accuracy shown in Table1. The average performance of these models is reported as about 69%, highlighting that, at least in this case, the cross-validation estimation of the general performance of the model was reasonable.

TABLE I. THE TRAINING RESULTS OF THE 5-FOLD METHOD

Vaid accuracy (%)					
Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*	*Mean*
71.05	65.79	73.68	65.79	67.76	68.82

The resulting models from the cross-validation process can be combined to provide a cross-validation ensemble that is likely to have better performance on an average than a given single model. Here all the 5 models are ensembled together. In the experiment, we are interacting over all the models to get the last layers as output. A merge layer will be added to compute the average output scores of all the models. The model is tested with 191 other new images. The accuracy achieved is 63.35% and the confusion matrix is shown in Fig. 7. Based on this confusion matrix with about 200 images being tested, there are 121 correct images and 70 incorrect images in classification. However, there is a major problem with the benign class. More than half of the benign images are classified on the wrong side.

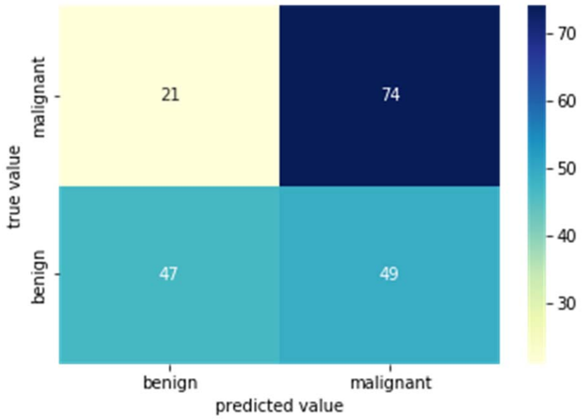


Fig. 7. The confusion matrix on 191 test images.

In Fig. 8 some samples of test images classified by the final model are depicted. The class of benign is labeled with green letters and malignant is labeled with red letters. Each of the images shows the prediction classes along with their probability.

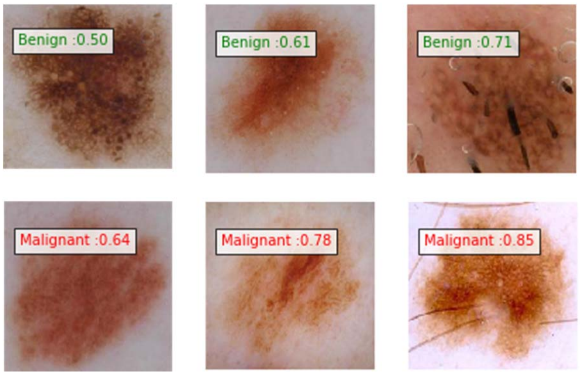


Fig. 8. Examples of images classified by the final model.

V. Conclusion and Future Work

The experiment was conducted on 760 images for training with 5-Fold cross validation method. It produced 5 models with 100 epochs resulting in a 63.35% accuracy of new test images.

The amount of training data and epoch used for training affects the level of accuracy in classifying benign and malignant images. To produce the best accuracy, it is important to set several proper parameters, such as optimal function, learning rate, the meta architecture and the resolution size of input images during training. The quality of the images and the correct annotation of images play a vital role, since it is crucial for the machine to learn the right information from human experience.

Deep melanoma diagnosis requires a large dataset due to the complex nature of the problem. In the case of melanomas, the discriminating features are rather weak, which makes the problem of high accuracy classification difficult, contrary to traditional image classification problems. There e.g., the identification of objects with marked characteristics are easy to understand even for the human eye from the object's features. Unfortunately, it is not easy to receive data from local hospitals in short time. In addition, the benign in this dataset, are difficult to get distinguished from malignant. In the future, researchers can train the machine to learn more images from benign and then try to make a comparison with different architectures on more melanoma cancer images.

References

[1] WHO (2019) Skin cancers - how common is the skin cancer? (World Health Organization (WHO)). Last accessed 15 May 2019.

[2] https://melanoma.canceraustralia.gov.au/statistics. Last accessed in 17 January 2020.

[3] A. Association, "2018 Alzheimer's disease facts and figures," Alzheimer's & Dementia 14.3 (2018): 367-429.

[4] M. E. Vestergaard, P. Macaskill, P. E. Holt, and S. W. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," British Journal of Dermatology, 159(3), pp. 669 – 676, Sep. 2008.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, Sep 4. 2014.

[6] A. GC. Pacheco, A. R. Ali and T. Trappenberg, "Skin cancer detection based on deep lea

[7] R. Refianti, A. B. Mutiara and R. Poetri Priyandini, "Classification of Melanoma Skin Cancer using Convolutional Neural Network," Int. J. of Advanced Computer Science and Applications 10.3 (2019).

[8] J. Jaworek-Korjakowska, "A Deep Learning Approach to Vascular Structure Segmentation in Dermoscopy Colour Images," BioMed research international 2018 (2018).

[9] R. D. Seeja and A. Suresh, "Deep Learning Based Skin Lesion Segmentation and Classification of Melanoma Using Support Vector Machine (SVM)," Asian Pacific Journal of Cancer Prevention: APJCP 20.5 (2019): 1555.

[10] N.C. Codella, Q.B. Nguyen, S. Pankanti, D.A. Gutman, B. Helba, A.C. Halpern and J.R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," IBM Journal of Research and Development 61.4/5 (2017): 5-1.

[11] W. Stolz, "ABCD rule of dermatoscopy—A new practical method for early recognition of malignant-melanoma," Eur. J. Dermatol., vol. 4, no. 7, pp. 521–527, 1994.

[12] X. Yuan, Z. Yang, G. Zouridakis, and N. Mullani, "SVM-based texture classification and application to early melanoma detection," in Proc. IEEE 28th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBS), Aug. 2006, pp. 4775–4778.

[13] J. F. Aitken, J. Pfitzner, D. Battistutta, P. K. O'Rourke, A. C. Green and N. G.Martin, "Reliability of computer image analysis of pigmented skin lesions of Australian adolescents," Cancer: Interdisciplinary International Journal of the American Cancer Society 78.2 (1996): 252-257.

[14] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems. 2012.

[15] http://www.dermoscopy.org/consensus/2d.asp. Last accessed in 19 January 2020.

[16] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu and J. Hu, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation." Computational Materials Science 171 (2020): 109203.

[17] https://machinelearningmastery.com/k-fold-cross-validation/. Last accessed in 19 January 2020.

[18] Y. LeCun, Y., Y. Bengio and G. Hinton, 2015, "Deep Learning", nature 521, no. 7553 (2015): 436-444.

[19] L. Deng and D. Yu, "Deep Learning: Methods and Applications", Foundations and Trends in Signal Processing, Vol. 7, No. 3 - 4, 2013.

[20] X. Zhang, J. Zhao and Y. LeCun, “Character-level convolutional networks for text classification,” Advances in neural information processing systems. 2015.

[21] N. Yalta, K. Nakadai and T. Ogata, “Sound source localization using deep learning models,” Journal of Robotics and Mechatronics 29.1 (2017): 37-48.

[22] J. A. A. Salido and C. Ruiz, “Using deep learning to detect melanoma in dermoscopy images,” Int J Mach Learn Comput 8.1 (2018): 61-68.

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

[24] F. Nunnari and D. Sonntag, “A CNN toolbox for skin cancer classification,” arXiv preprint arXiv:1908.08187 (2019).

[25] Skin Cancer Foundation, “Skin cancer facts & statistics,” (2019).

[26] G. Betta, G. Di Leo, G. Fabbrocini, A. Paolillo, P. Sommella, “Dermoscopic image-analysis system: Estimation of atypical pigment network and atypical vascular pattern”, IEEE International Workshop on Medical Measurement and Applications MeMeA 2006, pp. 63-67.

[27] G. Di Leo, A. Paolillo, P. Sommella, G. Fabbrocini, O. Rescigno, “A software tool for the diagnosis of melanomas automatic implementation of the 7-point check list method”, 2010 IEEE International Instrumentation and Measurement Technology Conference 12MTC 2010, pp. 886-891, 3–6 May 2010.

[28] G. Di Leo, G. Fabbrocini, A. Paolillo, P. Sommella, “ A web-based application for dermoscopic measurements and learning”, 2015 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2015; Torino; Italy; 7 May 2015 through 9 May 2015; pp. 279-284.

[29] T. Tieleman and G. Hinton, “RMSprop gradient optimization”, URL http://www. cs. toronto. edu/tijmen/csc321/slides/lecture_slides_lec6. pdf, 2014.