Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd
https://doi.org/10.53070/bbd.990574

Journal of Computer Science
ISSN: 2548-1304
Volume: IDAP-2021, Issue: Special, pp: 192-207
@Anatolian Science

Comparing Popular CNN Models for an Imbalanced Dataset of Dermoscopic Images

Erkan DUMAN*[1], Zafer TOLAN[2]

[1]Department of Computer Engineering, FIRAT University, Elazığ, Türkiye

[2]Department of Computer Engineering, FIRAT University, Elazığ, Türkiye

(erkanduman@firat.edu.tr, ztolan@hotmail.com)

Received: Sep.3, 2021 Accepted: Sep.16, 2021 Published: Oct.20, 2021

Abstract— In this study, the performance of popular convolution architectures against an imbalanced dataset is analyzed in detail with a multi-classing medical image processing application. Our selection for dermoscopic images is a large-scale and imbalanced dataset consisting of 10,015 colored lesion images belonging to 7 different skin diseases, was used as a benchmark. Images without pathological testing are labeled by specialist dermatologists who are members of International Skin Imaging Association. The f1-score was preferred as the measurement metric during the training phase of the convolution networks, which were trained with imbalanced dataset, and the area under the receiver operating characteristic curve and the confusion matrix of each model were calculated at the test phase. In the validation phase of convolution networks, k-fold cross validation technique was used. In addition, the filters obtained from the ImageNet dataset have been imported with the Transfer-Learning option. Fine-tuning was applied to the deepest convolution layers in order for these pre-trained models to develop themselves specific to our application. In order to prevent the overfit problem, the feature extraction outputs of the models were drop-out at a rate of 50% after flattening, and L2-regularization (weigh decay) was applied during the update phase of the weights. Although it is not the main purpose of the study, in order to partially improve the performance of convolution architectures, synthetic lesion images created with data-augmentation for the minor classes in the imbalanced dataset were included in the training process in a way that does not cause information leakage.

Keywords: *Popular CNN Models, Dermoscopic Images, Skin Diseases Diagnosis, Imbalanced Dataset*

1. INTRODUCTION

During the pandemic process we live in, the importance of physician decision support systems has become more evident. During this period, people hesitate to go to the hospital and try to get help from a distance as much as possible. Computer-aided systems used in categorizing colored lesion images, which are the subject of this study, are ideal examples for remote diagnosis [7, 18].

The success achieved in deep learning-based classification of colored lesion images has surpassed even dermatologists [4, 8]. In that study in 2019, the researchers revealed a classifier that can diagnose better than 136 of 157 dermatologists. In the study carried out in Germany; Dermatologists were asked to manually diagnose the possibility of to be melanoma disease for a lesion image as a value in the range of [0-1], whereas a convolution network trained with clinical images predicted better the same image [3]. In their studies, where the number of dermatologists participating from different institutes is quite high, it is very remarkable that machine estimates can make more accurate predictions than 85% of the

participants. This situation is a concrete example of the success point reached in convolution networks based diagnosis.

Convolution Neural Network (CNN) based image processing applications are at the forefront in the early diagnosis of skin cancers, the most common type of cancer [15, 16, 17]. As in all areas of deep learning, popular and assertive new architectures emerge every day in convolution models that process images [14, 19]. Even the oldest of the models for which performance comparison was made in this study is not older than 4 years old. The classification capabilities of popular CNN models on dermoscopic datasets discussed in this study were compared practically. How much success of which CNN model with how many parameters has been revealed?

Dermatoscopy is a widely used diagnostic technique to identify benign and malignant skin cancers [20, 21]. It is used to automatically diagnose skin diseases better than to examination with the naked eye. There are various hardware devices used for taking dermatoscopic images. Dermatoscopy, which can also be defined as the method of examining the skin with a microscope, is a method of computer-aided analysis of skin images [1]. With this method, the map of moles in the body is created and the localization of each point is determined. Then, dermatoscopic images are taken and recorded for each mole. Thus, it is possible to compare the new image to be obtained in the next control with the previous state by the specialist physician.

In this study, the largest international data set consisting of digital images needed by researchers who want to use artificial intelligence techniques in the diagnosis of skin diseases is offered as open source by the ISIC (International Skin Imaging Collaboration) community. For our lesion classification practice, we used the HAM10000 (Human Against Machine with 10.000 training images) dataset in the ISIC-Archive, which consists of colored skin images taken from dermoscopy devices [6, 28].

As with the HAM10000 dataset, in real-world applications, datasets are generally not uniformly distributed. Since the minority class in the datasets is a subset of rare examples, it is very difficult to try to find a balance by increasing the number of elements in this set. Although it is not the purpose of our study, we tried to partially increase the sample number of minor classes in the HAM10000 dataset with the data augmentation technique, albeit a little. The main purpose of our application; to analyze the inherent capabilities of popular CNN models on imbalanced dermoscopic image datasets. Performance improvement of each model can be achieved with synthetic image generation techniques [2, 29]. In our practice; In the HAM10000 dataset, it has been examined to what extent samples belonging to minor classes can be detected by popular CNN models. Because the worst case scenario of a medical test is to tell someone who is really sick that it is negative. This is very likely to happen in imbalanced datasets especially in very large ones such as HAM10000. In order to avoid this problem, it is possible to use popular CNN models whose basic characteristics are mentioned in the following paragraphs.

In this study, how susceptible popular CNN models are to dermoscopic images is examined in practice. GoogleNet, which Szegedy et al in 2014 [25] used as a team name in the ILSVRC competition, is the ancestor of Inception module-based architectures. It is a ConvNet family born with the basic idea of implementing Network-in-Network thinking on CNN layers. In this study, performance tests of V3 from Inception family and Xception architectures, which is a further developed version by Chollet [26, 5], have been performed. The Inception network family is one of the best performers on massive scale image databases such as both Imagenet and JFT [13]. The output of the Inception module, which consists of combining modules, each of which is a small network, in the form of parallel branches, is in the form of combining the outputs of the branches. Conv2D processes of 1x1 dimension are applied at the beginning of each branch and 3x3 in the continuation. Thanks to this structure, it becomes possible to discover separately spatial and channel-based qualities that are difficult or impossible to learn together.

Training very deep networks is a very difficult process. Because as you go deeper, the gradient values converge to zero, just like the fire that disappears into ashes. Layer architecture, which is named as the residual block and includes skip-connections or in other words, short-cuts, enables the gradient value to be transmitted to the depths without losing it [9]. In this way, very deep networks such as ResNet152, which we also use in our study, can be trained. As can be guessed from its name, this model, which consists of 152 layers, has been successfully trained with the gradient value [10]. In other words, as the depth of CNN models increases, the possibility of overfitting increases. The purpose of building very

deep networks such as ResNet152 is of course; It is not increasing the ability to overfit, but revealing new pattern discoveries with added layers. Thanks to the residual blocks in ResNet, the possibility of overfit is prevented. Thanks to the skip-connections within the residual blocks, it is ensured that the gradient value of the error in the output layer of the network reaches deep without getting lost. In addition, the bottleneck that may occur in any network layer is avoided, and the features learned before the layer are transferred beyond this bottleneck.

The idea of making parallel network structures in Inception modules can be trained quickly with short-cut connections in residual blocks led to the InceptionResNet architecture [24]. This hybrid construct is also called Inception V4. The training process has accelerated significantly compared to classical inception networks. In addition, the accuracy score value of the architecture created by assembling inception and residual blocks increases. While the Res-Net family tries to go deeper, the inception family tries to increase the number of parallelism in the horizontal. We can say that the InceptionResNetV2, for which we evaluate the performance in our study, is a hybrid architecture that has been optimized both deeply and transversely.

Developing and designing a CNN-based image classification application has now become almost an engineering field. The main idea in the birth of the NASNet architecture was how can we eliminate this need for design expertise [30]. The question of how we can teach a network to optimize itself by looking at its dataset is prone to costly answers, especially in large-scale real-world problems. Network Architecture Search idea; First, an optimum base block is built on a small-sized dataset, and then tens of these basic building blocks can be combined with their unique parameters and transferred to the large dataset. However, the construction of the basic building block and large building is similar to the problem of finding global minima in a search space and is very expensive.

Recent studies have shown that training in CNN networks can be more efficient if it contains shorter links between layers near the input and output. Dense Convolution Network (DenseNet) has the classic form of feed-forward connection but unlike it contains L*(L + 1) / 2 weights for L layers [12]. Each layer is connected to the layers after it by short circuits. In other words, the input of each layer is formed by the combination of the outputs of all the previous layers. There are 3 or 4 Dense Blocks in DenseNet and the transition between these blocks, that is, the size reduction, is performed with a transition layer. The transition called Transition Layer is conv + pooling process.

CNN-based intelligent image processing applications are the solutions needed not only in desktop or hosting systems, but also in hardware with limited resources such as smartphones that we can carry in our pocket. The number of parameters that can run on these limited resources is low, in other words, the need for lighter architectures has created the MobileNet family [11, 22]. The basic building block of this family are depthwise seperable convolution blocks. The depthwise seperable convolution block calculates one value for each channel, unlike the classical convolution process, and pasts them one after the other. Then, by applying 1x1 pointwise convolution to these outputs, the size of the model is very low. 1x1 pointwise convoluiton is the process of taking all the values in the same position in the filters and converting them into a single pixel. There are 2 extra hyperparameters such as latency and accuracy in the MobileNet family. These parameters allow the designer to determine the correct dimensions for his application.

Although a solution from the NASNet family costs much time and memory, its basic idea, that is, the idea of trying to optimize the architecture of the model according to its dataset has created a very important development. This new family, called EfficientNet [27], has chosen to scale the architecture with only one coefficient, instead of optimizing the architecture in a costly way, unlike NASNet. For example; When the current value of the scale is doubled, the depth in the architecture, the width and resolution of the intermediate outputs in the layer outputs also doubles. Although this new strategy, which can also be called combined scaling, is very simple but it gives very effective results. In our study, the achievement of the most effective results with the EfficientNetB5 architecture is an evidence of the efficiency of this strategy.

2. HAM10000 DATASET

In this study, the HAM10000 (Human Against Machine with 10.000 training images) data set consisting of 10,015 colored lesion images presented to the participants to train their models in the international competition organized by ISIC in 2018 was used as a benchmark [28]. The "ISIC Archive" is the largest global storage area where lesion images are made publicly available in digital image format. This dataset includes pre-processed dermoscopic images that can be used for early diagnosis of seven different skin diseases, along with Melanoma, the most widely known skin cancer.

These digital images taken from more than 10,000 patients in different countries using various devices and methods were used as a training data set at the international "Medical Image Computing and Computer Assisted Intervention - MICCAI 2018" conference held in Spain in 2018. In this contest, the models developed by the researchers reached the success point of nearly 90% correct disease diagnosis [6]. In ISIC competitions, the test data set is presented to the contestants with the real disease labels hidden in order to isolate the test data from the training data. For this purpose, in the HAM10000 dataset, besides 10,015 tagged training samples, 1,512 unlabeled lesion images were published. Because the ground-truth labels of the test data are not known and are not included in the training data; Overfitting of the classifiers developed by the competitors is prevented. Participants have been participating in the award-winning contest organized by ISIC since 2016, with short papers describing their algorithms and displaying their performance on the test data of the method they developed. In addition, in order for new researchers to test their performance, the website of the ISIC community has an online performance evaluation system for data sets separated by years (challenge.isic-archive.com). In this way, the success of each new algorithm developed, as in our study, on test data can be calculated instantly online.

Table 1-The descriptions and image counts of the all diseases in the HAM10000 dataset.

Disease	Name	Description	Count
MEL	Melanoma	It is a type of skin cancer that occurs with the uncontrolled division and proliferation of skin cells called melanocytes.	1.113
NV	Nevus	benign states of melanocytes.	6.705
BCC	Basal cell carcinoma	A type of cancer in basal cells that dies old cells in the skin and produces new ones instead.	514
AKIEC	Actinic keratosis	Prolonged sun exposure has a very low risk of developing cancer.	327
BKL	Benign keratosis	It is known as benign keratosis. They are generally seen in the elderly and are harmless skin growths.	1.099

DF	Dermatofibroma	It is a benign skin disease in terms of proliferation. They are hard, high-looking, fiber-like lesions that occur in different sizes.	115
VASC	Vascular lesion	Due to vascular problems, the superficial capillaries in the skin are enlarged or more frequent than normal.	142

In the HAM10000 dataset, besides the lesion images taken from deadly skin cancer patients such as "Melanoma (MEL)", it is good like "Nevus (NV)", which can be treated without surgical intervention, or "Actinic Keratosis (AKIEC)", which is often not even noticed because it is uncomfortable. There are also images of benign lesions. The abbreviations made for the names of the diseases in this dataset, the brief descriptions of the diseases and the total number of images for each disease are explained in Table 1. Throughout the study, the diseases were represented by abbreviations, just like the names of the columns in the HAM10000 dataset.

In multi-classing applications, positive samples belonging to the major class are much likely to be captured by the trained model. In other words; The model works more sensitive against the examples of the major class in the training class. However, as emphasized in Figure 1, if the HAM10000 dataset cake is cut according to the number of sample images of diseases; It is seen that there are 4 minority classes with a slice size of 5% or less. It is natural for the trained model to output biased against these small slices. Since classifiers are less sensitive to minority classes, the predictive output of classifiers trained with an imbalanced dataset is biased.

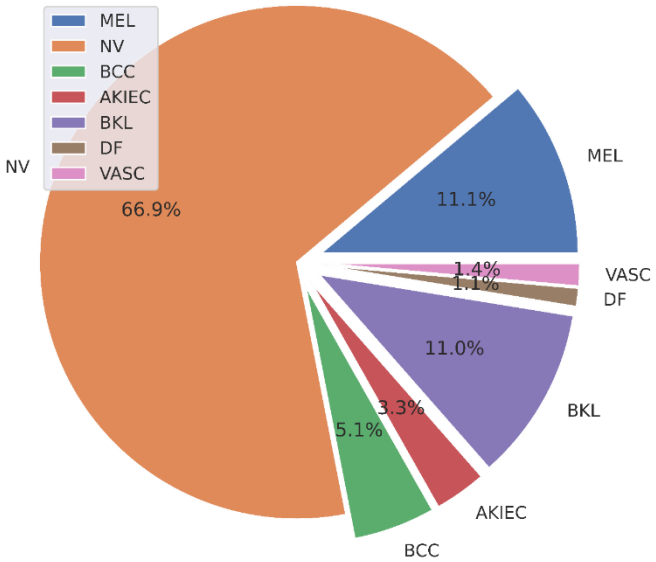

Figure 1. Percentage distribution of the number of images of the diseases in the HAM10000 dataset.

Unlike a binary classification problem, in multi-classing applications, it is inevitable that the number of negative samples is naturally high. Because in a classification problem with n outputs, the number of negative samples in the training data of any class is the sum of the positive samples belonging to the

remaining (n-1) classes. However, as shown in Figure 2, it is striking that the number of negative samples for all of the minor classes in the HAM10000 dataset is extremely high. In the HAM10000 dataset, the number of positive images of only major class "NV" disease is more than the number of negative images of the same disease. Therefore, positive sample numbers of all classes except NV class have been partially synthetically oversampled in this study. For the NV class, under-sampling is not preferred as it will result in data loss and make the features in the sample images of NV disease sacrificed from the classifier.

In healthcare, the worst case scenario is a false negative diagnosis (False Negative). In all physician decision support systems used, it is desired to identify as many cancer patients as possible to provide the care and treatment they need. In addition, the ability to diagnose the skin cancer diseases discussed in this article at an early stage greatly increases the probability of survival. For this reason, it would be a big mistake to select the accuracy value as metric when developing applications of datasets containing minority classes with a large number of negative samples such as HAM10000. Instead, a measurement that takes into account the recall, the False Negative rate, should be preferred. Accordingly, in our study, the f1-score score was used in the training and validation stages of the models. At the end of each epoch, f1-score values were recorded for both training and validation subsets of the dataset.

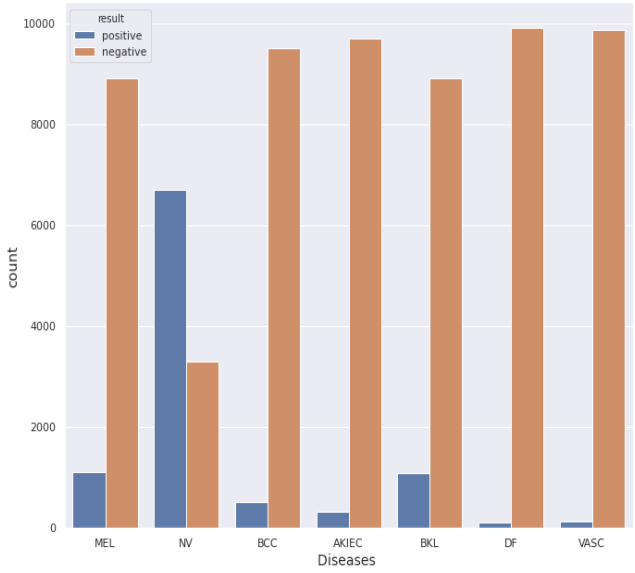

Figure 2. The Number of positive and negative labeled samples for each disease in HAM10000.

As can be remembered from the known equation in Equation 1, the f1-score score is the harmonic mean of the precision and recall metrics. The precision metric shows the probability that the person predicted as a patient in the output of the trained classifier is actually sick, while the recall metric shows what percentage of the samples that are actually sick can be detected. Recall metric is also called hit-rate. Since we want to punish our model for false-negative predictions as much as possible, the correct criteria for evaluating our model is the recall metric. In addition, in the off-line test dataset where the trained models are evaluated in local, the average of the recall values of all classes is also taken into consideration. Therefore, the performance of the models with high False Negative rate remained low. Classical optimization metrics such as accuracy cannot be used as an indicator of true performance, especially when the risk of false negative or false positive predictions increases.

$$f1\_score = 2 * \frac{precision * recall}{(precision + recall)}$$

$$precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \qquad Eq.1$$

$$recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

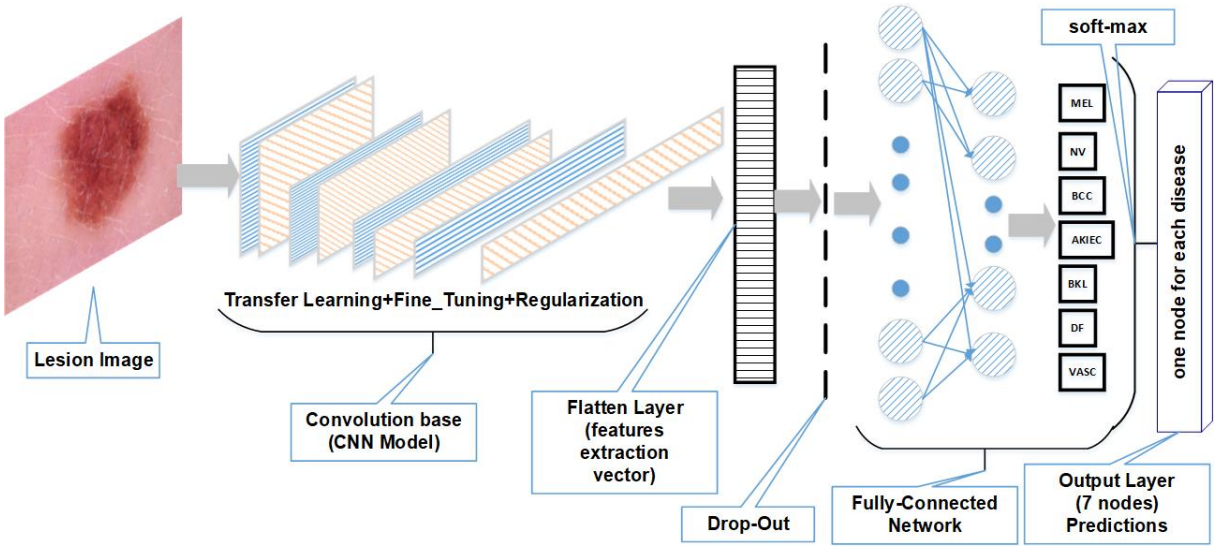

Figure 3.The basic diagram of adapted CNN models.

3. THE ADAPTATION AND TRAINING STEPS

The following adaptation steps were applied to all CNN models whose performance was compared against the HAM10000 dataset. The best possible performance has been tried to be obtained from each model considered. The original top classifier parts of all CNN models whose success has been tested have been eliminated. After the last convolution layer, the network output, named as feature extraction, has been applied flatten and the row has been turned into a vector. This feature vector obtained from CNN models is applied as an input to a 3-layer fully-connected Neural Network classifier as shown in Figure 3. This neural network classifier has 7 outputs, each corresponding to a disease. Therefore, when the lesion image is applied to the input of adapted CNN models shown in Figure 3, the probability of belonging to the related diseases at the specified 7 outputs [0-1] is obtained.

Step-1 (Transfer Learning): Starting with the filter values obtained from the classification of more than 14 million images into more than 21 thousand objects, that is, transferring the weight values trained with the Imagenet data set has been our basic starting point as in many studies. The components that are tried to be perceived in the initial convolution layers of CNN models have much simpler and generic features compared to the last layers.

Step-2 (Fine-Tuning): Only the convolution filters in the deepest layers of the CNN models were subjected to training. The filters used in the input layers were transferred from the imagenet data set and frozen. Both the training process is accelerated and the risk of overfit is reduced. The more parameters trained, the greater the risk of overfit. By fine-tuning the filters in the last layer of the CNN model and the subsequent fully-connected network, the examples in the HAM10000 dataset have been taught.

Step-3(Regularization): In order to prevent CNN models from overfitting, the Drop-out, L2-regularizaiton (weight-decay) and capacity reduction options were also considered. The last convolution layer outputs of the models were subjected to 50% drop-out after being flattened. As a second step for the overfit problem, the L2-regularization process was carried out by applying weigh decay (0.001) to the feature extraction part of the CNN models and the weight update steps of the Fully-Connected

classifier. In order for the recently developed models not to overfit, the number of layers and neurons in the Top-Classifier part was also reduced and the capacity was reduced.

Step-4(k-fold cross validation): Instead of classical hold-out validation, CNN models were trained and averaged with 5 different combinations of the training set. In this way, a possible error that could be made in the train-validation parsing of the training set was prevented. Of the 10.015 lesion images for training purposes in the HAM10000 dataset, 1.015 of them were used only for off-line (local) test purposes as shown in Figure 4. The local test set is unlikely to contain any samples from the training and validation sets. Likewise, every validation subset selected in each iteration is unlikely to contain the same image from the training subsets. In order to avoid any information leakage, 9.000 training images are separated into 5 equal parts in each iteration step and one of these parts, i.e. 1.800 images, is reserved for validation. A total of 7,200 pieces of images made up of the remaining 4 pieces are brought together and used for training the model. No data augmentation operation is performed for the validation subset selected in each iteration. Because if the data augmentation process is used for both training and validation, an information leak occurs. For this reason, the oversampling process was carried out after separating the training and validation clusters in each iteration. If k-fold and data augmentation are performed together without paying attention to information leakage, an overfitting problem can be observed that can increase the validation score value to 100%.

Step-5(Data Augmentation): In order to partially eliminate the imbalance in the HAM10000 dataset, every k. in iteration, positive samples of minor classes in training sets are synthetically replicated and doubled. There is no oversampling for only samples belonging to the NV major class. This reproduction is explained in detail in a numerical example given in Figure 4. For example, in any iteration; 9,000 off-line training images are primarily separated into 1,800 validating and 7,200 training. Then, the number of all the remaining lesion images, except NV disease images, in the 7,200 selected training images, is doubled with geometric transformation techniques. For example, in the numerical example in Figure 4, 700 images belonging to the MEL class were oversampled and increased to 1.400. Likewise, the sample numbers of all other minor classes were doubled and the total training set was transformed into 9.505 original + synthetic images from 7.200 original images.

During the Data Augmentation process, as shown in Figure 5, the number of training samples belonging to the minority cluster has been increased by activating the options of random rotation of the images, panning ratios of images horizontally and vertically, twisting process, image zooming, rotating the image vertically and filling the possible gaps on the new images. The other 3 new images in Figure 5 were obtained by applying random combinations of the above mentioned geometric processes to the original Melanoma image in the upper left corner of Figure 5.

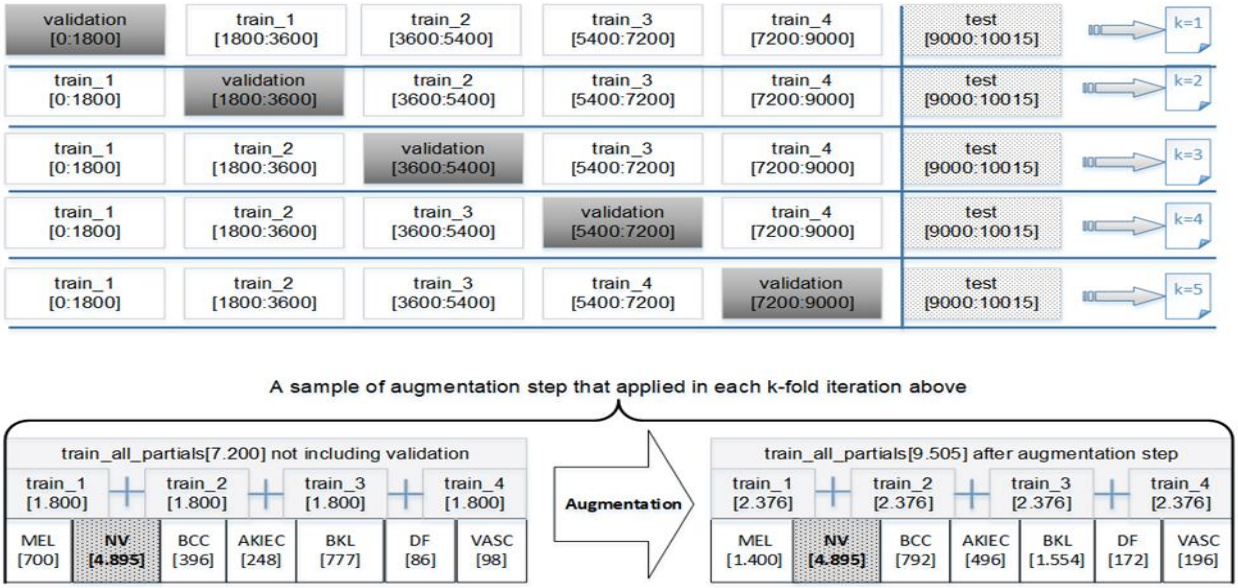

Figure 4. The data augmentation process that used in each iteration of k-fold cross validation.

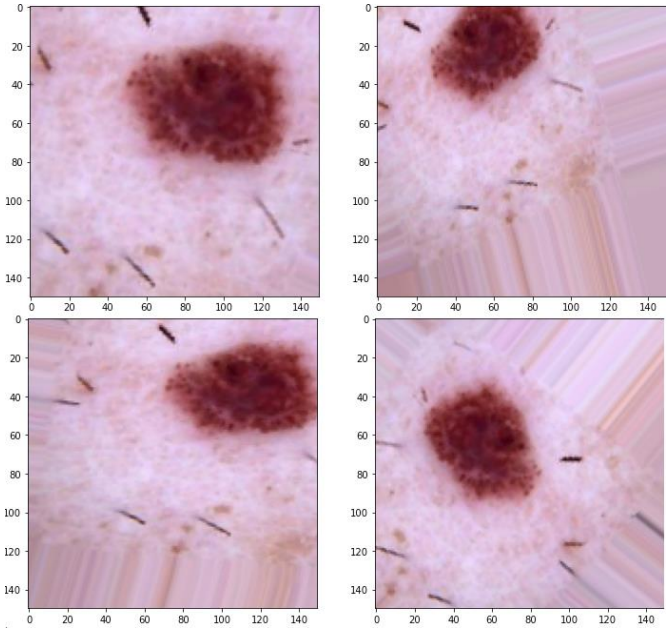

Figure 5. Generating synthetic images with Data Augmentation on a sample image of MEL disease

How the adaptation processes we applied in this study improve the success of a CNN model is shown in Figures 6 and 7. The f1-score graphs of the training and validation stages of a basic CNN model named VGG16, which can be considered as old [23], are given in Figure 6. Each adaptation step contributed positively to the success of the CNN model. Although the validation f1-score value did not come close to the one obtained from the training process, it is seen that the adaptation processes provide an improvement here as well. The pure CNN version in Figure 6 is the curve of the success achieved with only the transfer learning applied version of the VGG16 model. Adapted CNN version is the success curve obtained by applying fine-tuning, regularization and k-fold validation to the VGG16 model, respectively. Finally, the augmented CNN version is the success curve obtained as a result of all the adaptation processes described above, including synthetic data augmentation.

The ROC curve given in the upper part of Figure 7 and the zoom of this curve show the success of the predictions of the VGG16 models before the adaptation steps. Thanks to the adaptation processes, it is seen in Figure 7 that the accuracy of the predictions made by the VGG16 model after training has improved in the range of 15-20%. The AUC values of the Adapted VGG16 model appear to produce an average of 15% better results than in its pure state.

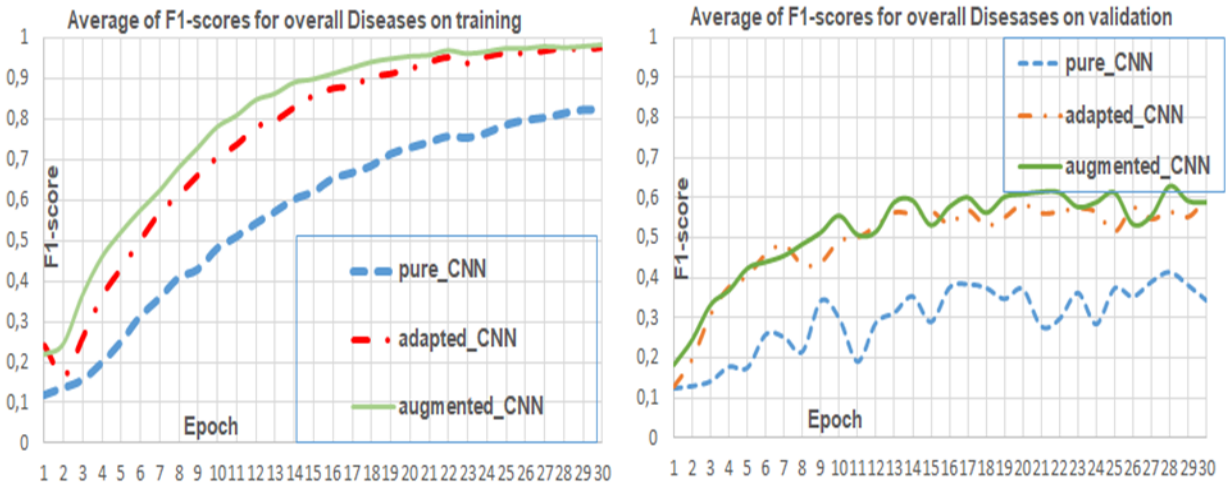

Figure 6. Comparison of f1-score values before and after the adaptation processes of the VGG16 model

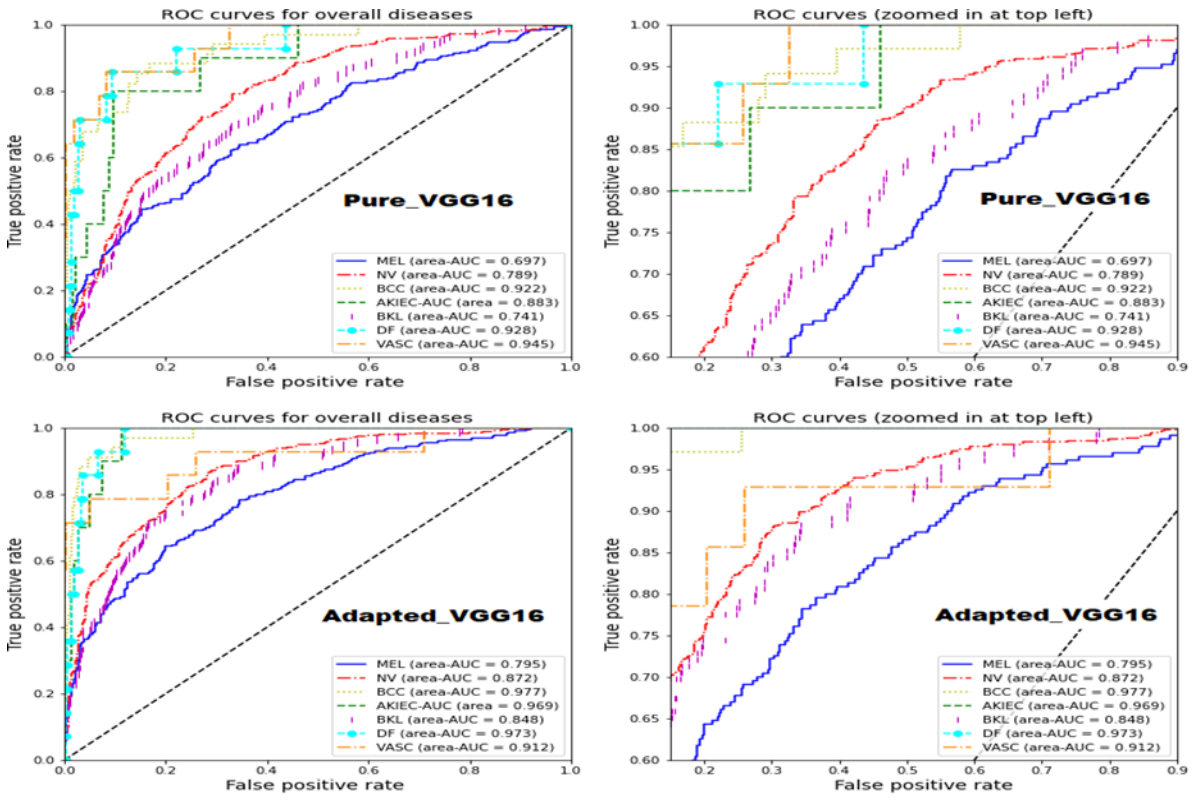

Figure 7. ROC curves of the predictions made by the VGG16 model for local-test data before and after the adaptation process.

4. EVALUATIONS OF CNN MODELS WITH OFFLINE-LOCAL TEST DATA

It was mentioned in Figure 4 that 1,015 training images were kept secret in order to perform off-line tests on the trained models. Since we have the actual label values of this test data, namely disease information, the estimates made for 1.015 images that were never seen by the models that completed the training and validation phase could be evaluated locally. In this way, the ROC curve showing how accurate the predictions of the trained model belonging to this off-line test set and the AUC values under it could be plotted separately for each disease class as shown in Figure 8.

The disease class MEL appears to be the most challenging of all models. This level of difficulty is the main reason for conducting research studies with only the Melanoma detection title. The CNN model with the highest AUC value of the MEL class has been the EfficientNetB5 architecture. EfficientNetB5 architecture stands out with its high AUC values for other disease classes as well.

Despite its small size, MobileNet architecture can make predictions with accuracy close to the EfficientNet family. Another result that should be mentioned is that the architecture with the lowest AUC values is NASNet. In fact, the optimization process of the model could have been long to improve the NASNet solution by paying the cost, but this would be an injustice for the other candidates.

It is seen that InceptionResNet architecture, which is the hybrid version of ResNet and Inception families, gives better results than its ancestors as expected. In addition, Xception, the close friend of this family, got ahead of both families.

It is seen that the DenseNet architecture, which was created by increasing the number of weights in the model inputs and outputs, also yields better results than the Inception family.

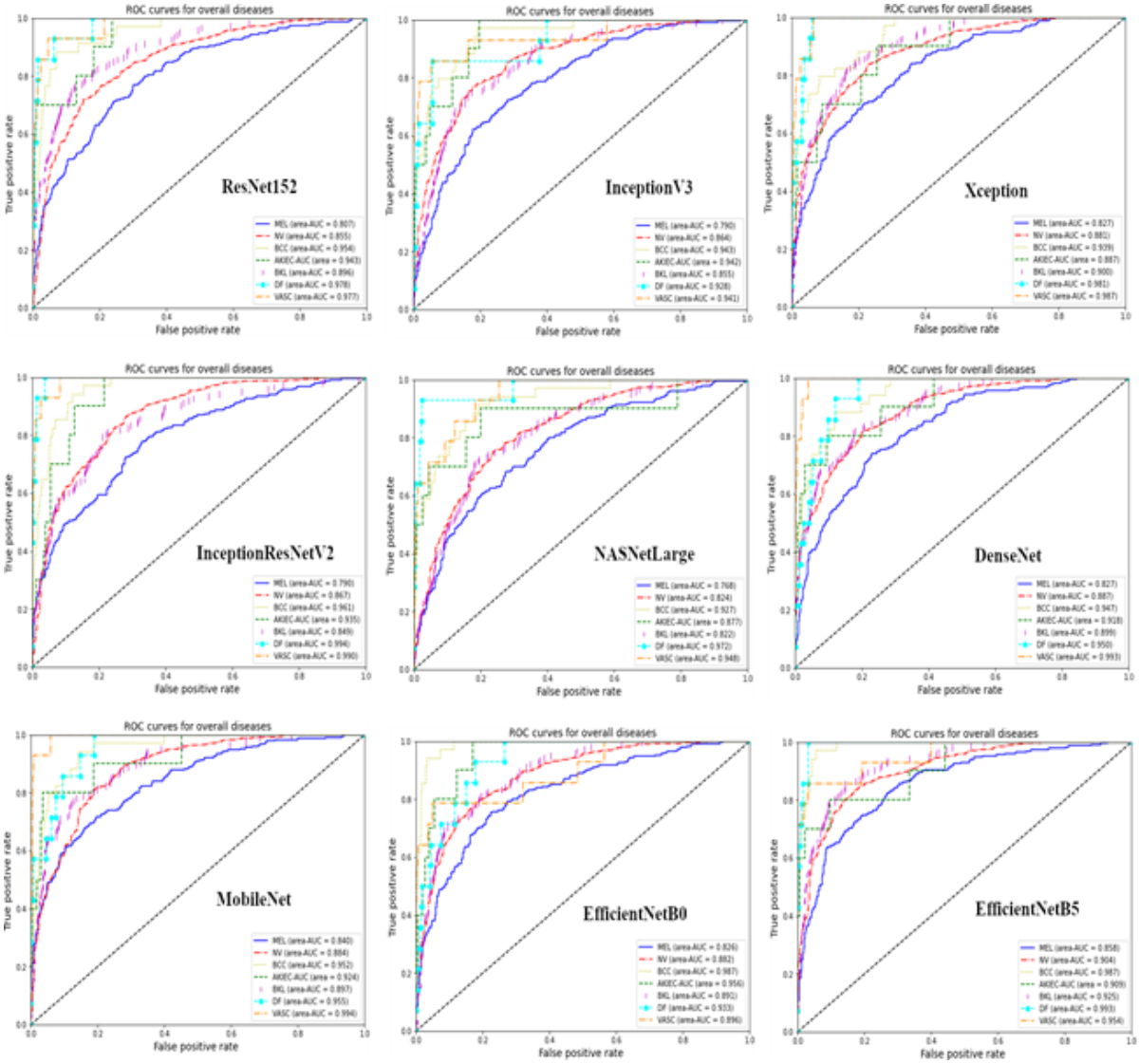

Figure 8. Comparison of ROC curves and AUC values of CNN models' predictions for test data.

The comparison of the predictions made by the CNN models for the off-line test data with the actual label values is shown in Figure 9. The sum of values in all cells in each of the confusion matrices in Figure 9 is 1.015. As in the distribution of the image numbers of diseases in the HAM10000 dataset, the darkest cell in confusion matrices belongs to the NV major class and is around 50% of the total of all cells. It is observed that the most successful predictions are made for NV, which is the major class in the HAM10000 dataset.

If we read a line from Figure 9; Consider, for example, the first line of the ResNet152 network. Only 107 of the 230 MEL dermoscopic images in the offline test set were correctly predicted. In fact, 74 of the images belonging to the MEL class were estimated as NV, 6 as BCC, 1 as AKIEC, 36 as BKL, 4 as DF and 2 as VASC. Thanks to confusion matrices, in a multi-classing application, it is revealed that the model confuses which disease with another or which one can predict with high accuracy.

In a multi-classing application, in the confusion matrix of the CNN model, cell values on the diagonal line are expected to be as high as possible. In the outputs we have obtained, it is seen that the closest CNN model to this situation is EfficientNetB5. The number of dark cells other than the diagonal cells indicates that the disease in the respective row is too confused with the disease in the respective column. The closest example of this bad scenario in our study is the NASNet architecture. However, it is an interesting point that the NASNet makes the highest correct estimate of the MEL class, which is the most difficult diagnosis for other models.

As can be seen in the ROC curves, the row with the highest number of false predictions of all CNN models is MEL disease. Another striking point is that the EfficientNetB5 architecture, the owner of the best predictions, MEL estimates remain lower than other models.

In addition, another important weakness should be emphasized that the sum of False Negative cells for all minor classes is close to the number in the True Positive Rate cell. This means that the predictions made remain around 50% accuracy on the basis of the disease. Of course it's not as bad as a coin flip, but it must be said that the error rate is high.

Figure 9. Comparison of confusion matrices of CNN models for off-line test data.

5. THE BATTLE OF CNN MODELS

Although we have isolated 1.015 images for off-line testing purposes, in fact, the HAM10000 dataset already has a real test set consisting of 1.512 images. However, since the tag values of this real test set are kept confidential, it is not possible to use it during the model development phase. Since the evaluation of this real test set is given only as Balanced Accuracy result on the on-line platform, as shown in Figure 10, in our study, 1.015 images in the training set were used as off-line test set in order to make a detailed evaluation locally. The ISIC online evaluation system has a weekly submission limit of 10. The Balanced Accuracy Score value is the average of the recall or hit-rate rates of the diseases.

The model, which was found successful in the off-line test process, was then subjected to the train process, starting from pre-trained weight values with the whole HAM10000 dataset, namely 10.015 lesion images. At this stage, it is no longer necessary to break down training data for validation or off-line testing. Performance metrics of CNN models trained with the whole dataset for the real test set were obtained from the online evaluation system of the ISIC as shown in Table-2.

Figure 11. The sample result mail including BAS-Score for an online submission

Table 2- Comparison of CNN Models' performance for Online-Real Test data and their parameters numbers.

CNN Model	Input Shape	Conv_base parameters	Required Memory	Trainable Parameters	Total parameters	Balanced Accuracy Score (offline test)	Balanced Accuracy Score (online test)
ResNet152V2	(224,224,3)	60,380,648	232 MB	117,9760,71	161,357,319	0.421	0.521
InceptionV3	(299,299,3)	23,851,784	92 MB	140,553,223	156,285,735	0.463	0.469
Xception	(299,299,3)	22,910,480	88 MB	216,756,087	230,841,903	0.473	0.507
InceptionResNetV2	(299,299,3)	55,873,736	215 MB	105,757,735	155,265,255	0.466	0.516
NASNetLarge	(331,331,3)	88,949,818	343 MB	135,892,039	216,189,529	0.438	0.437
DenseNet121	(224,224,3)	8,062,504	33 MB	53,775,367	58,682,951	0.486	0.518

MobileNet	(224,224,3)	4,253,864	16 MB	53,232,135	54,874,311	0.478	0.497
EfficientNetB0	(224,224,3)	5,330,571	29 MB	65,612,087	68,540,067	0.466	0.504
EfficientNetB5	(450,450,3)	30,562,527	118 MB	120,799,379	146,480,375	0.517	0.557

6.CONCLUSIONS

In Figure 11, the Balanced Accuracy Score values for offline and online test data sets are plotted and point sizes indicates the the number of parameters for related CNN models. Although the biggest point is the NASNet family, the accuracy of both off-line and on-line predictions remained lower than other families. In fact, the network architecture optimization process could have been made to perform better by keeping it too long. But it would mean spending all the working time just for this purpose; It would be a concession not given to other CNN models.

As it can be understood from the point representing the ResNet architecture, the distance between online and offline forecasts is greater than other convolution networks. The training with the complete dataset caused the performance increase of the ResNet family the most.

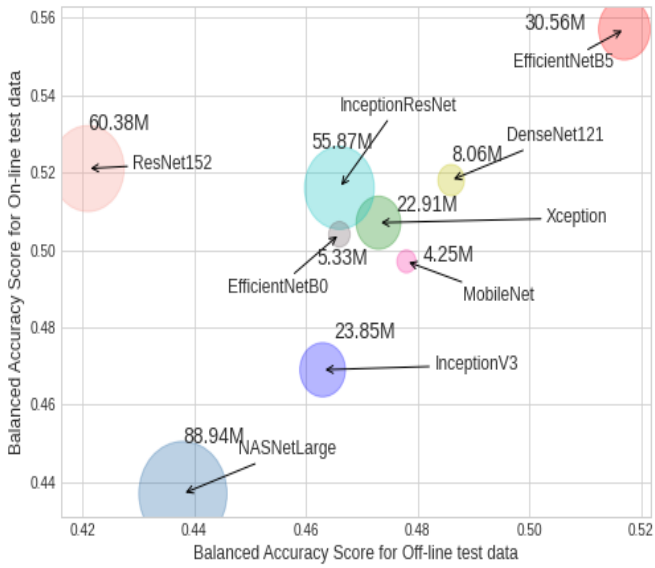

Figure. 11. Comparison of balanced accuracy score values and parameters numbers of CNN models.

Although the number of parameters is close, both the online and offline performance of the Xception point is better than its ancestor, the inception architecture. While InceptionResNet, whose parents are Inception and ResNet, performed much better than parents both offline and online evaluation. The InceptionResNet point has shown the benefit of the doubling of the Inception architecture as the number of parameters, with the increased accuracy of online and offline predictions.

Recently, efforts to achieve big performances with small network structures have highlighted themselves with the MobileNet point in Figure 11. Although the dot size is very small compared to other families, it surpasses many well-established families in predictive accuracy.

DenseNet architecture is also a point that shows high performance despite the small number of parameters. According to the 121-layer depth, the number of parameters is relatively small and the prediction accuracy is high.

When it comes to the leading architecture of the race, the Efficient family seems to have come to the fore with its 5th member, the EfficientNetB5 model. Since the dimensions of the images in the

HAM10000 dataset are 450x450, there is no need to measure the performance of the 6 and 7 convolution networks of this family. A significant performance increase of 10% is observed between the B0 model and the B5 model.

References

Barata, C., Celebi, M. E., & Marques, J. S. (2019, 5). A Survey of Feature Extraction in Dermoscopy Image Analysis of Skin Cancer. IEEE Journal of Biomedical and Health Informatics, 23, 1096–1109. doi:10.1109/jbhi.2018.2845939

Bisla, D., Choromanska, A., Berman, R. S., Stein, J. A., & Polsky, D. (2019). Towards Automated Melanoma Detection With Deep Learning: Data Purification and Augmentation. (pp. 2720–2728). Long Beach, CA, USA: IEEE. doi:10.1109/CVPRW.2019.00330

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., . . . Schrüfer, P. (2019, 4). A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. European Journal of Cancer, 111, 148–154. doi:10.1016/j.ejca.2019.02.005

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., . . . Schrüfer, P. (2019, 5). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer, 113, 47–54. doi:10.1016/j.ejca.2019.04.001 Chollet, F. (2016, 10). Xception: Deep Learning with Depthwise Separable Convolutions.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., . . . Halpern, A. (2019, 2). Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC).

Do, T. T., Hoang, T., Pomponiu, V., Zhou, Y., Chen, Z., Cheung, N. M., . . . Tan, S. H. (2017, 11). Accessible Melanoma Detection using Smartphones and Mobile Image Analysis.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017, 1). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115–118. doi:10.1038/nature21056

He, K., Zhang, X., Ren, S., & Sun, J. (2015, 12). Deep Residual Learning for Image Recognition.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, 3). Identity Mappings in Deep Residual Networks.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017, 4). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016, 8). Densely Connected Convolutional Networks.

Ioffe, S., & Szegedy, C. (2015, 2). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

Kassani, S. H., & Kassani, P. H. (2019, 6). A comparative study of deep learning architectures on melanoma detection. Tissue and Cell, 58, 76–83. doi:10.1016/j.tice.2019.04.009

Li, Y., & Shen, L. (2018, 2). Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. Sensors, 18, 556. doi:10.3390/s18020556

Majtner, T., Yildirim-Yayilgan, S., & Hardeberg, J. Y. (2018, 10). Optimised deep learning features for improved melanoma detection. Multimedia Tools and Applications, 78, 11883–11903. doi:10.1007/s11042-018-6734-6

Mishra, N. K., & Celebi, M. E. (2016, 1). An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning.

Okuboyejo, D. A., & Olugbara, O. O. (2018, 4). A Review of Prevalent Methods for Automatic Skin Lesion Diagnosis. The Open Dermatology Journal, 12, 14–53. doi:10.2174/187437220181201014

Rezvantalab, A., Safigholi, H., & Karimijeshni, S. (2018, 10). Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms.

Rosendahl, C., McColl, A. C., & Wilkinson, D. (2012, 7). Dermatoscopy in routine practice - 'chaos and clues'. australian family physician, 41(7), 482-487.

Salido, J. A., & Jr., C. R. (2018, 2). Using Deep Learning for Melanoma Detection in Dermoscopy Images. International Journal of Machine Learning and Computing, 8, 61–68. doi:10.18178/ijmlc.2018.8.1.664

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018, 1). MobileNetV2: Inverted Residuals and Linear Bottlenecks. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520.

Simonyan, K., & Zisserman, A. (2014, 9). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016, 2). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2014, 9). Going Deeper with Convolutions.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015, 12). Rethinking the Inception Architecture for Computer Vision.

Tan, M., & Le, Q. V. (2019, 5). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. International Conference on Machine Learning, 2019.

Tschandl, P., Rosendahl, C., & Kittler, H. (2018, 3). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 (2018). doi:10.1038/sdata.2018.161

Vasconcelos, C. N., & Vasconcelos, B. N. (2017, 2). Convolutional Neural Network Committees for Melanoma Classification with Classical And Expert Knowledge Based Image Transforms Data Augmentation.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2017, 7). Learning Transferable Architectures for Scalable Image Recognition.