Skin Anomaly Detection Using Classification Algorithms

A.D. Andronescu, D.I. Nastac
Faculty of Electronics, Telecommunications and Information Technology POLITEHNICA University of Bucharest, Romania
nastac@ieee.org

G.S. Tiplica
Colentina Hospital at Carol Davila University of Medicine and Pharmacy, Bucharest, Romania

Abstract—**This paper proposes a solution for the most easily detectable type of cancer, skin cancer. Due to the ozone layer thinning, it was observed an increase in the number of skin lesion anomalies. Those anomalies are produced by the increasing UV radiation intensity. Good protection against sun is not always facile, mainly in the exposed regions of the skin like the face and the lower regions of the arms. Most types of cancers can be easily extracted via surgery if they are detected in one of the incipient development stages. Using a high resolution image, a classification algorithm can detect certain patterns in the nevi color, shape, veins or irregularities inside. Analyzing those aspects, a person can track their nevi development easily and intervene if it necessary. The originality of this paper consist in designing several convolutional neural networks, using the Python programming language, the Keras API and the Tensorflow framework alongside with proper dataset selection. The convolutional neural networks were trained using an open source dataset, with the purpose of formulating a diagnosis for new patients. The convolutional neural networks are provided with 7 classes of images, representing 7 types of skin lesions as input, and will output a diagnosis. The result of the research was a classifier with a convenient classification accuracy based on the dataset, and a medium accuracy based on new data.**

Keywords—skin cancer, nevi, classification algorithm, artificial neural network

I. Introduction

The cancer problem is considered one of the biggest challenges of the 21st century. The incidence of cancers is measured to be about 400 cases / 100.000 people / year, and the number of mortality cases exceeds 160 / 100.000 people / year [1]. The mortality rate could easily decrease if the cancers would be detected in an incipient stage when an surgical operation would remove the tumor. Many cancer detection methods are invasive and require special medical equipment, but the skin cancer can be detected with the naked eye. The dermatoscope is, in theory an instrument that helps the dermatologist diagnose pigmented lesions, having a magnifying glass and a light source. In terms of the light source used, two types of dermatoscopes exist: those who use polarized light in order to observe the deeper layers of the epidermis and non polarized light in order to observe the surface of the lesion. A photo instrument (like a camera of a mobile phone) can capture a photo that can suffice certain requirements in terms of quality if it has a certain resolution and illumination source, as a non polarized dermatoscope.

The aim of paper is to investigate on an automatic approach for skin cancer diagnosis. This task will be performed using a deep learning approach, using a classification algorithm. The starting point of this project was the article published in the Nature magazine: ”The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”, published: 14 August 2018, in the Nature magazine [2]. In order to build from scratch a CNN, some initial parameters needed to be analyzed. In this manner, to observe the influence of the parameters, a tool called “NVIDIA” digits [4] was used in order to train fast using different models: LeNet [5], Alex-Net [6], Google Net [7]. Using the results obtained from the pre-trained networks, parameters in terms images were chosen.

II. The Dataset

The dataset represents one of the most important requirements when building a deep learning algorithm and a convolutional neural network. Choosing a dataset however is not an easy task. The dataset needs to contain many images (the more images, the better). The complexity of a dataset is directly proportional with the complexity of the machine learning algorithm. Also, one should expect to obtain a lower performance since the classification error will increase proportionally with the number of classes, but inversely proportional with the size of the training dataset.

As stated before, this paper was inspired by the ”The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”, which introduced an open-source dataset used for skin diseases as described in [3]. This dataset was collected in a period of 20 years of dermatoscopic images gathering, resulting over 10000 images, where more than 50% of them were confirmed through histopathologic diagnoses. The literature on this domain (skin cancer detection), however do not cover the entire range of skin lesions because most of them are focusing only on one or few types of cancer: melanoma. In this paper, we tried to extend to a wider class of skin lesions.

A. Data selection and organization

The input data selection is considered an important step in building a classification model. The dataset consists of 7 image classes: Actinic Keratoses and Intraepithelial Carcinoma (Bowen's disease) - 327 images, Basal Cell Calcinoma - 514 images, Benign keratosis- 1099 images, Dermatofibroma - 115 images, Melanoma - 1113 images, Melanocytic nevi - 6705 images, Vascular skin lesions - 142 images.

As observed, there is a category that contains more than 65% of the images in the dataset. This dataset does contain unbalanced categories, in terms of training data for out model. Consequently, this could produce a huge impact on the performance of the algorithm. To explain this, let us assume that the classification algorithm will classify all images as "nevi", keeping in mind that the nevi image set consists over 65% of all images. This will imply a 65% accuracy. Considering that the algorithm has to identify correctly the category out of 7 types, one would say that the probability is of 1/7 (14%, in the case of a balanced set of data). Comparing with about 65% accuracy obtained for an unbalanced set of data, the result is obviously useless. A better accuracy metric can be considered by comparing the number of "false positive" classifications with the number of "false negative" ones. In the above stated case, the number of "false positive" classifications would be 0, whereas the number of "false negative" classifications would be about 2000 cases. In medicine, this is considered as unacceptable since "false negative" cases could be malignant cancers that might lead to serious health issues of the patient, or even death. Since some categories had very few images, the algorithm cannot predict them very well, and they will be neglected. In order to counteract this problem, data augmentation should be performed. This preprocessing part represents the action of enlarging the data available for training and testing. This step is performed in order to improve the relevance of the results, counteract the problem of over-fitting and increase the detection probability of the classes having a small number of images. Since the dataset already contains regions of interest selection, data augmentation by zooming and cropping were discarded. Data augmentation by rotation was used in order to induce a balance in the number of image classes. Thus, augmentation by rotation was used in order to increase the number of images up to 500 images per class for the classes containing less than 500 images. For the classes containing more than 500 images, no augmentation was performed, and only 500 images were selected. Thus, the training and testing dataset was brought to 3500 images, containing balanced classes that would produce better results than using the whole dataset. However, having a smaller number of images used, will result also to smaller classification accuracy, and a smaller ability to generalize.

B. *Dataset format*

When building a neural network, a few parameters need to be set, parameters that dictate how to further build the neural network. The parameters that we found crucial on the dataset were: the size of the input image, the type of image (greyscale / RGB). In order to set those parameters from the beginning, the problem was approached using pre-trained networks, in order to observe the influence of those factors.

NVIDIA Deep Learning GPU Training System (DIGITS) [4] is a NVCaffe and Tensorflow wrapper that provides a graphical web interface to the previously mentioned frameworks.

This tool can be used to rapidly train deep neural networks suitable for image classification, segmentation, image detection. It also simplifies the task of data management and speeds up the training process, providing already implemented CUDA compatibility. Since this tool already provides 3 state-of the art, pre-trained deep neural networks that use different image parameters, it was used in order to decide what type of parameters were to be used .

TABLE I. PERFORMANCE OF PRE-TRAINED NEURAL NETWORKS

Model	**Image resolution**	**Image type**	**Results**		
			Validation Accuracy	***Loss (Validation)***	***Loss (training)***
LeNet [5]	28x28	Greyscale	70.98%	0.86	0.83
Alex-Net[6]	256x256	Greyscale	69.80 %	0.84	0.92
Google-Net[7]	256x256	RGB	71.43%	0.73	0.59

Analyzing the results, the accuracies were similar. We found out that a 28x28 pixel image will provide decent resolution for the classification system. This is the case when images were acquired in a special environment (hospitals), using special a special tool (dermatoscope), a process which selected the region of interest, having always similar lighting conditions. For this reason, we decided to use images with higher resolution (256x256 pixels) in order to extend the extend the capabilities of using the classification model, with regular photo cameras. Also, we observed that the LeNet architecture reached a plateau (limit), where improvements were not presented over a period of 5 epochs. This is why we rejected the use of low-resolution images in the implementation of the project. After comparing the results for Google Net and Alex Net models, we observed that Google Net architecture was producing better results, but only by a small margin. Since Google Net architecture was trained using RGB images, we considered its performance as a standard for networks working with RGB images. We observed that the improvement brought by RGB images would not suffice for the computational resources required. Also, using the actual colors for the diagnostic, may be inadequate since exposure, saturation of colors, color temperature and white balance differ from photo device to photo device. In dermatology, not necessarily the color influences the diagnostic, but the number of colors that are present on a specific nevi, aspect that can be detected even with greyscale imaging. By choosing an RGB image in detriment of a greyscale image, we would require 3 times more computational power.

III. THE NEURAL NETWORK

The Convolutional Neural Network that had the best results consisted in a model with 2 convolutional layers. The convolutional layers were designed using the Keras API.

One important aspect that the Keras API is bringing to the models is a random weights initialization. Because of this aspect, one should not expect to obtain exactly the same results after several trainings. In order to obtain better results using the same model, more training processes should be performed.

The model with the highest accuracy, obtained a 72.14% accuracy. This model was composed of 2 convolutional layers consisting in 64, respectively 128 filters with 3x3 and 5x5 padding sizes. In order to reduce the dimensions of the feature map resulted, Max Pooling was used with 2x2 and 5x5 padding size. The feature map was flatten, then fed to 2 fully connected layers, containing 128 neurons each.

Overfitting represents the main problem of the neural networks, consequently two overfitting control techniques were used:

- Dropout layers
- Regularizers

Dropout layers with 25% and 30% hiding ratio were used before each fully connected layer.

Regularizers are functions that have the main purpose of reducing the overfitting and increasing the generalization error. In the used environment, 2 types of regularizers exist: L1 and L2. L1 regularization technique is called Lasso Regression [8] and L2 is called Ridge Regression [9]. The Lasso Regression implies neglecting the coefficients that are close to 0, thus losing information, and reducing the training time. It is a good method of feature selection. The Ridge Regression involves adding the squared magnitude of the coefficients to the loss function. In the given model, the lamda coefficient in the kernel regularizer has been set to 0.01, as a result of trial and error.

The architecture of the model that provided the best result is presented in figure 1.

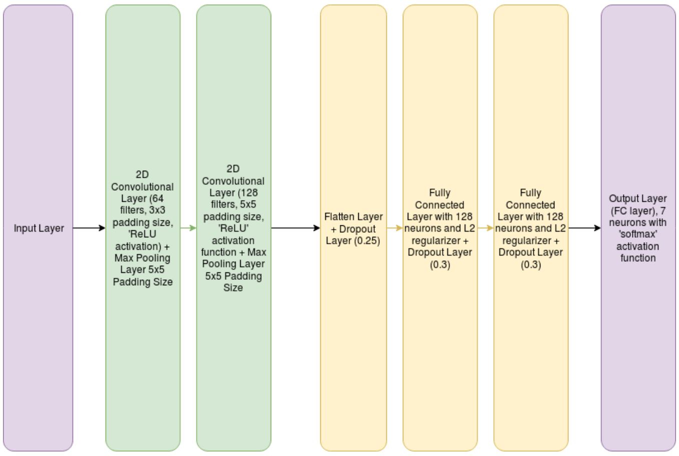
Input Layer
2D Convolutional Layer (64 filters, 3x3 padding size, 'ReLU activation) + Max Pooling Layer 5x5 Padding Size
2D Convolutional Layer (128 filters, 5x5 padding size, 'ReLU' activation function + Max Pooling Layer 5x5 Padding Size
Flatten Layer + Dropout Layer (0.25)
Fully Connected Layer with 128 neurons and L2 regularizer + Dropout Layer (0.3)
Fully Connected Layer with 128 neurons and L2 regularizer + Dropout Layer (0.3)
Output Layer (FC layer), 7 neurons with 'softmax' activation function

Fig. 1. The CNN architecture

In order to improve the performance of the network, certain functions have been used in order to combat the overfitting problem. After a certain point, the model will not learn from the training data, and its validation accuracy will decrease.

In order to fight this behavior, a certain set of function is used. This set of functions are called ”Callbacks”, and they can be applied at different stages of the training procedure. Two ”Callback” functions have been used in this training set:

- Early stopping
- Model save

The early stopping procedure is a procedure used to stop the training session if no improvements are observed after a given number of epochs. The Model save is a function that will ensure that only the model that has given the highest accuracy.

IV. Results

A 72.14% validation accuracy was obtained and the training accuracy was 93.17%. The validation loss was also different from the : validation loss was 2.0401 and the training loss: 1.086. The fact that the training accuracy was not close to 100% is in general a good indicator that the model has some ability to generalize. However, overfitting, which is a negative event, still occurred starting from the 10th epoch, as shown in the figure 2 and 3.

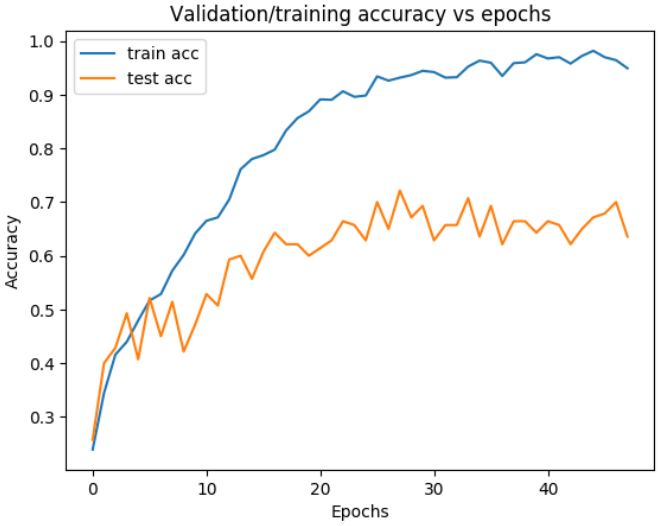

Fig. 2 The evolution of the accuracy during the training and testing phases

In order to observe how good the resulted classification function really is, the loss function is calculated, as we can easily observe in figure 3.

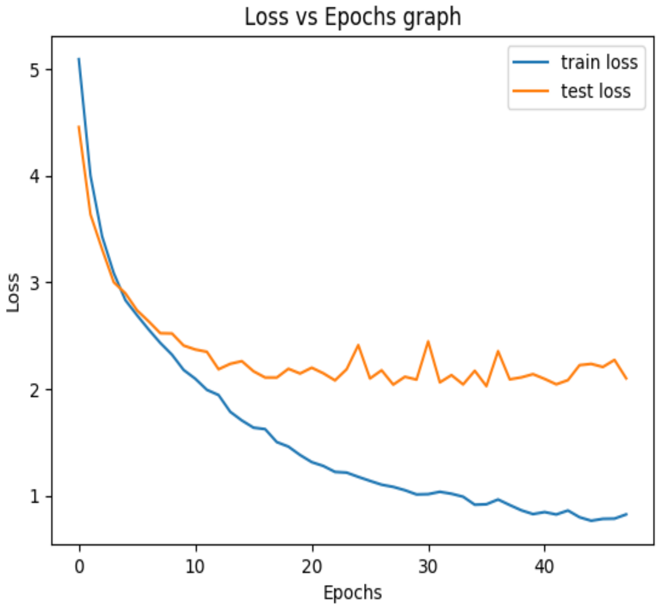

Fig. 3 The evolution of the loss function during the training and testing phases

By observing the figures 2 and 3, it can clearly observed that the overfitting event occurs after a certain point. This is one of the most difficult tasks to overcome in machine learning. However, a 72% accuracy is satisfactory for a 7 classes classifier. Transposed in real terms, it indicates that in almost ¾ cases, the algorithm will correctly classify the image to the correct class. If given a random guess, the classifying accuracy would be of 14%. More models have been tested, and the best results will be further presented.

As a starting point, a model containing 2 hidden layers (2 convolutional layers) have been used, and tuned in order to increase its testing accuracy, and reduce the training time. The effect of different regularizers on the model will be further presented in Table II.

TABLE II. THE EFFECT OF REGULARIZERS

Regularizer	Training accuracy	Test Accuracy	*Loss (Test)*	*Loss (Training)*
L2	93.17%	72.14%	2.04	1.08
L1	68.6%	57.9%	1.99	1.62
L1 & L2	36.2%	42.02%	6.61	6.61

On behalf of the algorithm hyperparameters and optimizers, testes have been executed in order to choose the most appropriate optimization algorithm. The Adam optimizer is an extension to the classical stochastic gradient decent. In the table III, a comparison between the results obtained by the Adam optimizer and stochastic gradient descent is presented. The Adam optimizer used a relatively small learning rate that will decay as the training process is occurring, allowing for a fine-tuning of the parameters of the network.

TABLE III. THE EFFECT OF ALGORITHM OPTIMIZERS ON THE MODEL

Optimizer	Training accuracy	Test Accuracy	*Loss (Test)*	*Loss (Training)*
Adam (lr = 1e-4,decay = 1e-6	93.17%	72.14%	2.04	1.08
stochastic gradient descent	97.22%	57.9%	5.62	4.30

As observed, using the simple stochastic gradient descent optimizer in the training algorithm, only accentuates the overfitting event.

Different numbers of convolutional layers have been used in order to achieve better results. Considering the same training process was used, the best accuracy obtained by a change in the number of feature extraction layers (convolutional layers) will be presented in the Table IV.

TABLE IV. THE EFFECT OF NUMBER OF HIDDEN LAYERS

Number of convolutional layers	Training accuracy	Test Accuracy	*Loss (Test)*	*Loss (Training)*
2	93.17%	72.14%	2.04	1.08
3	96.13%	68.38%	2.5	0.99
7	99.18%	60.2%	1.55	0.001

As observed, having a model with more hidden layers does solve the problem of high loss as expected. The model is able to learn more features. However the increase in layers number exponentially increases the training time. Another problem occurs when models with a high number of hidden layers, the overfitting problem is harder to control.

New data has been tested using new images that have been acquired from the Dermatological Department of the Colentina Hospital. Since the dataset used for training the model has been formed from images from the same medium, using a reduced number of dermatological equipment, the performance of the network would only be relevant if it can used can be used with images provided from different dermatological equipment . Using diagnosis provided by 3 dermatology specialists, the following results were obtained (Table V):

TABLE V. DOCTOR PROVIDED DIAGNOSIS

Detected lesion	Total number of images classified	Correct	*Wrong*
Actinic Keratoses	12	4	8
Benign Keratosis	19	9	10
Dermatofibroma	4	3	1 (melanoma)
Melanoma	19	11	8
Vasculary lesions	2	0	2 (melanoma)
Melanocytic nevi	5	5	0
Basal Cell Calcinoma	0	0	0

In medicine, the least desired outcome is represented by the false negative diagnosis. In order to compute the number of false positive and false negative diagnostics, a few assumptions have been made. The classes that were considered malignant were: melanoma, actinic keratoses, basal cell carcinoma. Even though actinic keratoses, and basal cell carcinoma have a chance smaller than 20% of becoming malignant, all of them were considered as dangerous and a consult would be required if those would be detected. The classes: melanocytic nevi, benign keratosis, vasculary lesions, dermatofibrom are known to be benign cancers or skin lesions and were not considered harmful to the human body.

From the table V, the following results were deduced (table VI):

TABLE VI. PERFORMANCE OF THE NETWORK ON NEW DATA

False Negative	False Positive	Accuracy
3 / 61	18 / 61	52%

V. DISCUSSION

In order to train the algorithm, the HAM10000 dataset has been used, provided by [3]. In order to maximize the efficiency of the training, data augmentation by rotation has been performed to increase the number of categories that presented a small number of images. Out of the 10015 original images and 1914 augmented images, 3500 images were selected, 500 images. The number of 500 images per class has been chosen due to the unbalanced nature of the number of image categories. In essence, the augmentation by rotation does not introduce new information in the system; it was only used as an artificial mean of dataset balancing. In addition, in order to increase the entropy of the dataset, images that introduced new information in the system (non augmented images), were always chosen in detriment of augmented ones. Out of the 3500 images, 80% have been used for algorithm training and 20% for validation. In addition, a test dataset has been provided by the Colentina Hospital in order to test the performance of the model. An accuracy of 72.14% has been achieved on the validation set, and 93.56% accuracy for the training set. Even if the model is overfitted, the validation accuracy shows that this model is still usable, and has some ability to generalize to an extent. The most important parameter in the paper that should be minimized as much as possible is still the number of false negatives. From table VI, it can be observed that a chance of giving a false negative diagnostic is smaller than 5%.

Hyperparameters tuning has been performed in order to obtain acceptable validation accuracy. Tests have been performed on this behalf:

- Used more convolutional layers
- Used variation of convolutional layers parameters (32, 64,128 filters, different padding sizes)
- Used dropout layers with different parameters
- Used different optimizers
- Used Model Checkpoints and Early Stopping procedures
- Used different learning rates
- Used different regularization methods

The chosen configuration was a simple one, consisting in 2 convolutional layers, 2 dropout layers, placed before and between 2 fully connected layers, a small batch size (32), and a small learning rate: 1e-4 with decay. The most efficient from the validation accuracy point of view was the presented one. Variations of this model or other models were tested, but they did not reach over 71% validation accuracy, even though the models with more convolutional layers had a better training accuracy.

Further improvements of the project can include:

- Model implementation using Decision Trees in order to achieve a better accuracy ;
- Using RGB images with a higher resolution;
- Involving different data augmentation techniques (adding Gaussian noise, scaling);
- Gather more data, either by merging the available datasets available, or by gathering new data by visiting hospitals, etc. ;
- Insert heuristics in order to differentiate better between nevi and melanoma based on the ABCDE rule [10] ;
- Testing more architectures, with more layers that provide a lower loss.

VI. Conclusions

In conclusion, our paper highlights the benefits of engineering in the medical context, by improving the quality of life. It stands as a solution to the flooding of information, information that can be much faster be processed by a machine in comparison to a human being. The amount of information currently available overwhelms every human being, and deep learning is a solution to make use of them. However, one challenges that are encountered are the need for information classification, which will allow the machines to learn. In the medical context, this tool can provide doctors and patients with a complementary diagnose tool.

In this paper, a solution for predicting the skin lesions has been presented, by taking advantage of deep learning and convolutional neural networks. The image classifier was developed by implementing the structure of a convolutional neural network. The final model was developed through trial and error techniques and variation of the algorithm and model hyperparameters. This paper proves that skin lesions can be classified using this approach even though improvements can still be performed. Our personal contribution to the project consists in processing the data in order to achieve relevant results by means of image selection and augmentation. We designed and tested networks having a small number of hidden layers selecting the highest accuracy ones, tuned them in order to increase the prediction accuracy. The relevance of the results was tested using specialized support from Colentina dermato-venereology Department, who provided us with new data and advised medical expertise. The results, based on new data, do provide high accuracy in terms of false-negative classifications. However improvements can be made in order to increase the actual class-based classification. One crucial factor that needs to be kept in mind is that a certain diagnostic cannot be provided based solely upon images prided in a single time instance. In order to accurately improve the accuracy, the evolution in time of the disease should be also monitored. Other factors impact the lesion prognosis, like: skin color, region, usage of protection barriers (sunscreen, adequate clothing), the UV radiation exposure (tanning beds, periods of high solar activity), age, sex, etc. The results obtained by implementing this prediction algorithm represents just another proof of the capability of machine learning and its ability to adapt and generalize when faced with uncharted territories.

References

[1] Cancer Statistics, published by the National Cancer Institute , April 27, 2018, https://www.cancer.gov/about-cancer/understanding/statistics

[2] Cliff Rosendahl Harald Kittler Philipp Tschandl , “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”, published in naturesearch journal – scientific data (2018)

[3] Philipp Tschandl,2018 , The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Harvard Dataverse, https://doi.org/10.7910/DVN/DBW86T

[4] NVIDIA Corporation. Deep learning digits documentation. https://docs.nvidia.com/deeplearning/digits/digits-user-guide/index.html, 2018. accessed on August, 2019.

[5] Yann LeCun, Leon Bottou, Yoshima Bengio and Patrick Haffner, 1998, - Gradient Based Learning Applied to Document Recognition, Proc of the IEEE, November 1998 available: http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf

[6] *Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton - ImageNet Classification with Deep ConvolutionalNeural Networks, published in Communications of the ACM,Volume 60 Issue 6, June 2017 Pages 84-90*

[7] *Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich - Going deeper with convolutions, published in 2014*

[8] *Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88.*

[9] *Ng, Andrew Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance, published in ICML ‘04 Proceedings of the twenty-first international conference on Machine learning Page 78.*

[10] *Naheed R. Abbasi, MPH, MD; Helen M. Shaw, PhD; Darrell S. Rigel, Robert J. Friedman, MD; William H. McCarthy, FRACS; Iman Osman, MD; A;fred W. Kopf, MD; David Poloski, MD, PhD . Early Diagnosis of Cutaneous Melanoma, published in NCBI, 2004*