SKIN LESION CLASSIFICATION: TRANSFORMATION-BASED APPROACH TO CONVOLUTIONAL NEURAL NETWORKS

Charles Lehman [1], Martin Halicek[2,3], Ghassan Alregib[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology
[2]Department of Biomedical Engineering, Georgia Tech & Emory University
[3]Medical College of Georgia, Augusta, GA

ABSTRACT

Diagnosing malignant skin lesions early is often the difference between life or death. With the increasing accessibility of deep learning tools that have demonstrated outstanding performance for image classification, it is no surprise that there has been an extensive effort to employ neural networks in the diagnosis of skin lesions. We explore a method of late-fusion of three identical CNN's models, trained with three different image transformations (RGB, FFT, and HSV) of the same dataset. The resulting fused accuracy of 98% is a 4% increase to each lone network.

***Index Terms*—** Neural Network, Convolutional Neural Network, Melanoma, Classification

1. INTRODUCTION

1.1. Pigmented Skin Lesion Diagnosis

Pigmented skin lesions are neoplastic growths of melanin-producing cells, called melanocytes. Melanin is the predominant pigment in human skin and hair, which is responsible for variations in skin tone. The uncontrolled growth of this type of tissue can be divided into either benign (relatively slow growing and non-cancerous) or malignant (invasive, with potential for metastasis to other tissues), which are called nevi and melanoma, respectively. There are two other types of cancer that develop on the skin, squamous cell carcinoma and basal cell carcinoma, which are developed from different cell types. Melanoma has a much higher risk of metastasis than these two types and has a much higher mortality, so it was chosen as the target for this investigation because early detection greatly improves survival [1].

Patients with suspicious skin lesions are typically referred to a skin specialist doctor, a dermatologist. The initial diagnosis of melanoma is made by visual inspection using the following acronym. We will refer to this criteria as the ABCDE criteria. **A** - Asymmetry (benign lesions tend to be more circular and symmetrical), **B** - Border (benign lesions have well-defined borders, malignant spread and taper), **C** - Color (malignant lesions typically have multiple hues of color), **D** - Diameter (malignant lesions are typically larger than 6mm), and **E** - Evolution (increased rate of change in any of the above factors, or onset of clinical symptoms are indicative of malignancy).

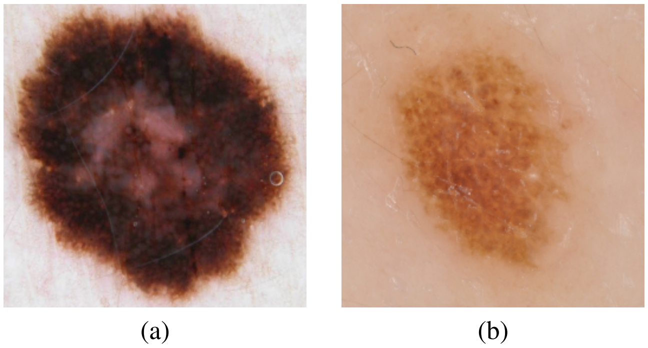


Fig. 1. Representative benign (a) and malignant (b) lesion from the ISIC Archive [2].

1.2. Prior Art

Early detection of melanoma is crucial, so the area has been researched extensively in attempt to automatically diagnose malignant lesions. To our knowledge, the first application of artificial neural networks to diagnose melanoma was presented in [3]. They reported an accuracy of 80% on two-class (malignant or benign) classification [3].

Similar works have reproduced these results. More recently, [4] used 2-dimensional wavelet transforms of pigmented skin lesions for feature extraction to train a neural network. The authors reported 100% sensitivity and specificity, but the reproducibility of their work is limited, as they used only 31 images for training and 21 images for testing. However, this indicates an expansion of the field into more computationally intensive methods.

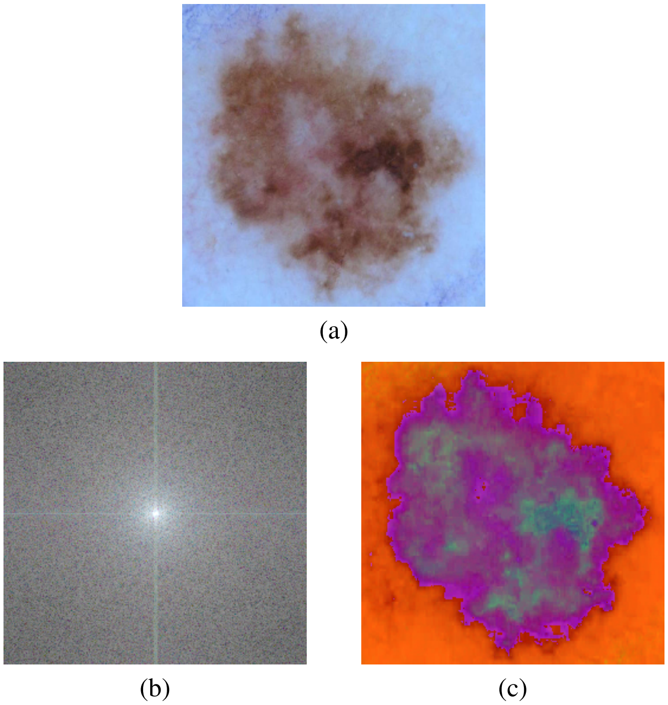
Two recent and relevant works incorporate support vector machines and neural networks on large contemporary data sets. [5] uses a non-dermoscopic data set of 1300 images, including lesions in 10 categories of approximately equal sample sizes [5]. They employ the deep neural network, *AlexNet* [6], and test with leave-one-out cross-validation. The comparative result for their two-class (malignant or benign) accuracy for their technique was 94.8%. However, using only an average of 130 images for each of the 10 categories could lead to data over-fitting and limited reproducibilty of the results.

Another group, [7], uses the same data set presented here, the `ISIC-Archive`, but they employ SVM, deep learning, and sparse coding classifications. Their comparative results for two-class accuracy range from 85.3-91.9% for their Caffe CNNs, and a reported 93.1% accuracy when fusing Caffe and sparse coding features.

2. METHOD

2.1. Skin Lesion Dataset Processing

The International Skin Imaging Collaboration: Melanoma Project (ISIC) provides a database of digital images of pigmented skin lesions with associated metadata for use in the study of "digitally-assisted melanoma diagnosis." [2] The `ISIC-Archive` was selected as the source for our dataset due to the large number of images that have been diagnosed by experts, which functions as the ground truth classification. Also, the availability of metadata, specifically, diagnosis and subclassifications provides the necessary information to generate a multi-layer classification problem.



(a)

(b)

(c)

Fig. 2. Example transformations of a pigmented skin lesion from RGB (a) [2] to FFT (b) and HSV (c).

The images were cropped prior to or in lieu of downsampling to produce an area that contained the most entire part of the lesion. This helps keep the dataset scale-invariant as many of the benign lesions were relatively small compared to the malignant lesions. Also, many benign lesion images contained a circular colored sticker, which were not present in the malignant lesion images. The cropping was done manually and duplicate images or images containing too little information were excluded (i.e. skin lesion that comprised less than 5% of total pixels in the image).

```
< label >< Ch.1 >< Ch.2 >< Ch.3 >
                                ...
< label >< Ch.1 >< Ch.2 >< Ch.3 >
```

Fig. 3. Structure of binary files [8]

Images that were too large or too small were resampled into 256px×256px squares using bicubic interpolation as provided in OpenCV [9]. This process induced distortions to the small subset of images that were not square. This operation could affect visual methods of diagnosis utilizing ABCDE criteria. Also, there was no consideration on differentiating effects on predictions between downsampling and upsampling.

It is important to note that resampling was conducted before any transformations as the 256px×256px is the basis of comparison for the two transformations. Individual color channels within each Red, Green, Blue (RGB) image were transformed via Fast Fourier Transform (FFT) and Hue, Saturation, Luminance (HSV) mappings to create three separate transformations of identical datasets, as depicted in Fig. 4. Human-Interpretable Structure (HIS) is necessary to convey comparisons of the images that undergo transformations in the context of useful information. HIS can be considered spatiotemporal interpretations of an image (i.e. shapes, textures, colors, position, motion, etc.) as performed by a human. [10] Transformation of images from RGB to FFT seemingly loses HIS, while RGB to HSV maintains some HIS in the form of high-contrast segmentation as demonstrated in Fig. 2. This is not to say that an RGB to FFT transformation will lose all HIS, as high-frequency content (i.e. high-contrast lines) is easily discernible within the 2-D Fourier Domain. If quantization error is ignored, both transformations can be reversed without loss of information. This further suggests that useful information is observer dependent and that transformations can be utilized to optimize observations of useful information.

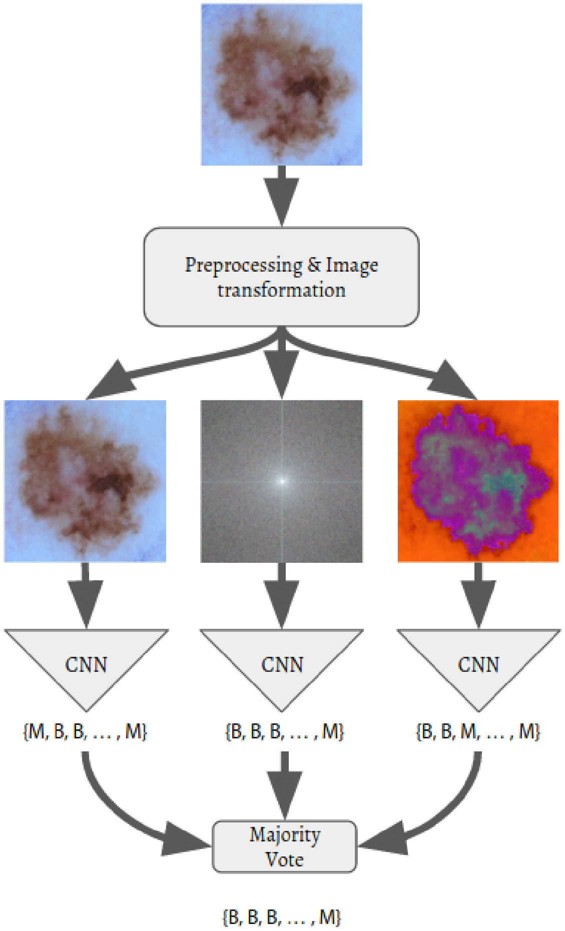


Fig. 4. Example transformations of a pigmented skin lesion from RGB (a) to FFT (b) and HSV (c).

The transformed datasets were converted to a format suitable for the CIFAR-10 model [11] provided with TensorFlow [12] (Fig. 3). We developed tools in Python to aid in handling batch conversion to binary files containing 300 256px × 256px, 3-channel images with their corresponding labels. There were 14 binaries generated per transformation and each binary was 57 M in size. The order of images and labels were identical across the transformations.

2.2. Skin Lesion Classifier

The CNN TensorFlow tutorial example model used to train a classifier for the CIFAR-10 dataset was utilized in this experiment. This model is meant to demonstrate the power of CNNs and the ease of their implementation [12]. We chose this model because the TensorFlow environment required very little modification to software in order to implement our experiment. This minimized the amount of development time while preserving the structure of the experiment.

2.2.1. Model Nascent State

Original image size = 256
Cropped image size = 220
Number of examples per epoch for training = 10,000
Moving Average Decay = 0.999
Number of epochs per decay = 350.0
Learning rate decay factor = 0.03
Initial learning rate = 0.03
Batch size = 8 (10 for evaluation)
Model weights and biases left as-is to establish nascent point.

The hardware used for the experiment was a ZOTAC MAGNUS EN760 mini-pc with an NVIDIA GeForce GTX-960 GPU. The GPU was utilized to run the CNN.

2.2.2. Training and Evaluation

Each instance of the CNN was trained for 28,000 steps which took approximately 4 hours. 2,700 images were used for the purpose of training and 1,500 images for evaluation. These two sets were independent and did not share images. The number of images for training was artificially increased through random crops, random blurring, and random rotations. Also, images were selected at random to generate batches for training. [8] [12]. Images used for evaluation were not altered in this way and were selected in order.

3. RESULTS

3.1. Evaluation Setup

Each instance of the CNN would perform inferencing on 1,500 images and output two vectors of data. The first would contain the labels of each image and the second would provide logits of correct classification. The image order across transformations was the same.

3.2. Fusion Method

The output inferences of the three CNNs were fused via majority vote. Figures 5 and 7 were generated from the outputs of the three CNNs. There is sufficient variation in consensus that results in an improved performance with majority vote.

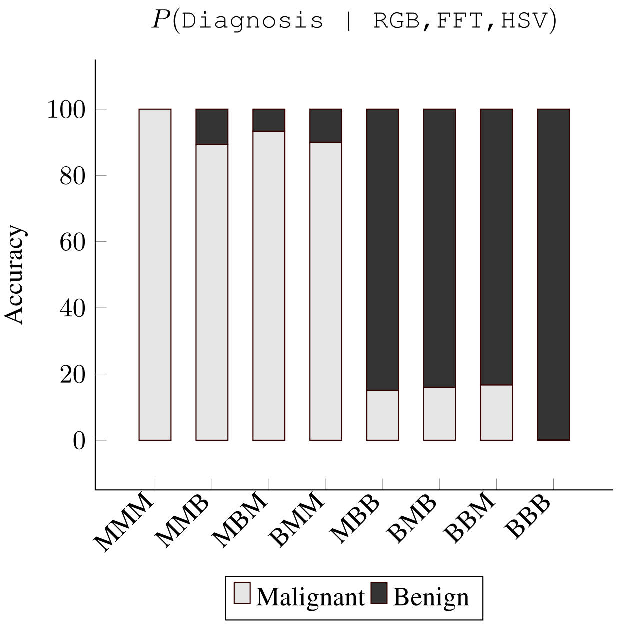


Fig. 5. Probability distribution of diagnosis given prediction vote, which is indicated by an M for malignant and B for benign 3 tuple MBM for RGB - M, FFT - B, and HSV - M.

3.3. Statistical Analysis

The statistical analysis methods used in the this paper are standard in medical diagnostic studies. Figure 2 are confusion matrices that break down true-positives (TP), false-positives (FP), false-negatives (FN), and true-negatives (TN).

3.4. Performance

	Accuracy
[5] Kawahara *et al.*, 2016	94.8%
[7] Codella *et al.*	93.1%
Ours	**97.6%**

Fig. 6. Accuracy of comparable techniques

Each of the three CNN's performed at about 93% accuracy individually. This improved to 97.6% after fusion. This result is possible because the inference streams are varying on their output sufficiently to yield improvement in performance. More interestingly, there was a improvement across all statistics (Fig 9). This result has maintained consistence with repeated training and evaluation cycles with variations to image order with each binary file and distribution of benign and malignant images.

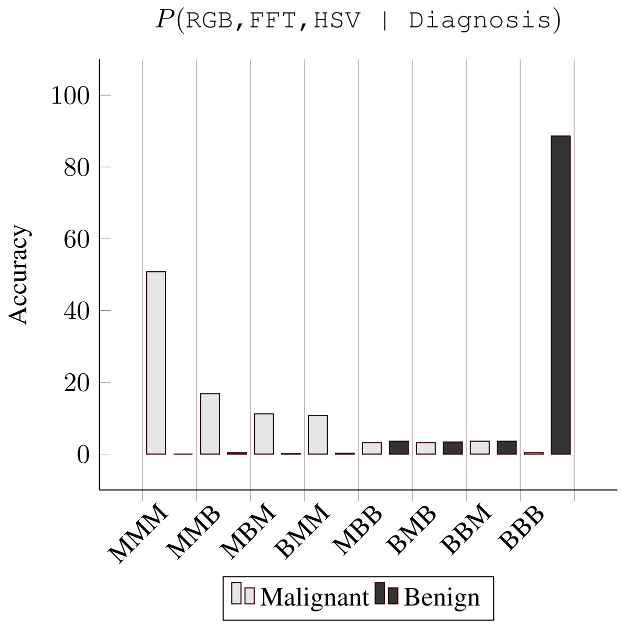


Fig. 7. Probability distribution of predictions given diagnosis, which is indicated by an M for malignant and B for benign 3 tuple MBM for RGB - M, FFT - B, and HSV - M.

		Predicted	
		Malignant	Benign
Actual	Malignant	205	45
	Benign	52	1198

(a) RGB

		Predicted	
		Malignant	Benign
Actual	Malignant	204	46
	Benign	50	1200

(b) FFT

		Predicted	
		Malignant	Benign
Actual	Malignant	191	59
	Benign	50	1200

(c) HSV

		Predicted	
		Malignant	Benign
Actual	Malignant	224	26
	Benign	10	1240

(d) Fused

Fig. 8. Results of 1500 image evaluation on tranformation-based CNN after 28,000 training steps

4. CONCLUSION

The performance of the CNN's individually were comparable to that of similar techniques. Also, we observed that there

	RGB	FFT	HSV	Fused
Sensitivity	82.00%	81.60%	76.40%	89.60%
Specificity	95.84%	96.00%	96.00%	99.20%
Precision (PPV)	79.77%	80.31%	79.25%	95.73%
NPV	96.38%	96.31%	95.31%	97.95%
Accuracy	93.53%	93.60%	92.73%	97.60%

Fig. 9. Performance of transformation-based CNN

is sufficient useful information within each transformation to achieve similar performance. Moreover, there is a significant jump in performance using the rather rudimentary fusion method of majority vote. This suggests that there is a variation on useful information between transformations sufficient enough to be used as additional evidence when making inferences. The observations made in this experiment are encouraging and they pave the way for new applications that span a broad range of signals analysis. Nevertheless, the plan is to repeat these experiments with considerably larger data sets and more than two classes. Further investigation into classifying transformations in terms of useful information is necessary to appropriately describe this phenomenon.

5. REFERENCES

[1] J. Sosman, "Patient education: Melanoma treatment; localized melanoma (beyond the basics)," 2016, In: UpToDate, Atkins MB (Ed), UpToDate, Waltham, MA. (Accessed 12 Nov. 2016).

[2] C. Curiel *et al.*, "International skin imaging collaboration: Melanoma project," 2016, Dataset available from isic-archive.com.

[3] F. Ercal *et al.*, "Neural network diagnosis of malignant melanoma from color images," *IEEE Trans Biomed Eng*, vol. 41, no. 9, pp. 837–845, Sep 1994.

[4] J. Abdul, S. Salim, and R. Aswin, "Artificial neural network based skin cancer detection," *International Journal of Advanced Research in Electrical*, vol. 1, pp. 200–5, 2012.

[5] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, April 2016, pp. 1397–1400.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[7] N. Codella and *et al.*, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," *18th International Conference on Medical Image Computing and Computer Assisted Intervention*, vol. 118, Jan 2015.

[8] A. Krizhevsky, "Cifar-10 dataset," 2009, Dataset Available from www.cs.toronto.edu/ kriz/cifar.html.

[9] G. Bradski, "Opencv library," *Dr. Dobb's Journal of Software Tools*, 2000, Documentation available at docs.opencv.org/3.1.0/.

[10] A. Pentland, "Perceptual organization and the representation of natural form," *Artificial Intelligence*, vol. 28, no. 3, pp. 293–331, 1986.

[11] A. Krizhevsky, "Convolutional deep belief networks on cifar-10," 2010.

[12] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.