DermFollow

A System For Better Diagnosis and Treatment of Skin Cancer

Matt May
Georgia Institute of Technology
maym@gatech.edu

Thanh Dang
Georgia Institute of Technology
thanhdang@gatech.edu

Stefano Fenu
Georgia Institute of Technology
sfenu3@gatech.edu

Apurv Verma
Georgia Institute of Technology
apurvverma@gatech.edu

Matt Cimino
Georgia Institute of Technology
ciminomc@gatech.edu

1. INTRODUCTION/MOTIVATION

The current standard-of-care for diagnosing skin lesions (such as melanoma) is in-person care in a dermatologist's office. A dermatologist performs a physical exam to review the patient for probable cancerous lesions. However, follow-up with the patient when he or she is outside of the office is minimal, resulting in unmonitored growth of potentially cancerous lesions over time.

In this paper we describe DermFollow, an application we have built to solve this problem. Patients use their computer or smartphone to upload pictures of their lesions over time. The system uses the knowledge learned from several thousand images via deep neural networks to compute the risk score for the uploaded image and present the most high-risk images, and therefore high-risk patients, to the clinician.

For each uploaded image, we also present the most similar images from the training set to the provider, to explain why (or why not) the network considers an image as a high-risk case. The clinician can use this added information and schedule an appointment with these patients on a priority basis.

2. PROBLEM DEFINITION

Our goal is to build a web application that promotes more effective diagnosis and treatment of skin cancer through machine learning-based analysis of patients' skin lesions over time.

Our system enables this through these key innovations: (1) risk analysis (classification) of suspected skin cancer lesions using deep residual and convolutional neural networks (CNNs), (2) dynamic patient risk scoring based on demographics, medical history, and patient image analysis, and (3) an interactive body map feature that encourages physician-patient communication. We hope to improve patient outcomes, raise patient satisfaction, and improve the efficiency of medical practice.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

3. SURVEY

The current state of the art for automated skin cancer analysis generally involves feature detection [12], [7], [14]. In dermatology, there is an ABCDE algorithm [21], [13], [18] that is used by doctors to assess potential skin cancer lesions. Existing approaches often attempt to imitate this algorithm via detection of the same features [22]. Some of these models have yielded poor results on the more diverse images encountered in practice, limiting adoption.

Neural network-based approaches have also been used for skin cancer classification [16], [5], [2], [3]. Kreutz et al. [10] use neural networks and feature extraction to classify skin lesions. Sheha et al. [15] use a multilayer perceptron to classify melanoma, attaining 92% accuracy. Esteva et al. [6] use an ensemble of CNNs to attain 90% binary classification (malignant/benign). While these results are impressive, many of these models were trained on images that lack histological (microscopic) verification, the gold standard for determining malignancy.

4. PROPOSED METHOD

4.1 Intuition

Low physician and patient engagement, limited integration into clinical workflows, poor algorithm performance on actual patient images, and expensive hardware requirements have plagued existing approaches.

To increase provider and patient engagement, DermFollow presents an interactive user interface, leveraging JavaScript d3 [1] to allow the physician to review a *body map* of the patient that displays the patient's images superimposed over the relevant point on the body. DermFollow also allows physicians to instantly send short messages to the patient after he or she uploads an image, giving the patient instant feedback.

DermFollow integrates seamlessly with the largely electronic clinical workflow, as it is Web-based and operates on the desktop and mobile devices.

To improve algorithm performance, we use an ensemble of deep CNNs and residual neural networks, which have been shown to be very effective at image classification [11], [20], [17]. Lastly, we use an image dataset with 100% histological verification.

4.2 Description

4.2.1 *User Interface*

The application has two interfaces, namely the patient interface and provider interface.

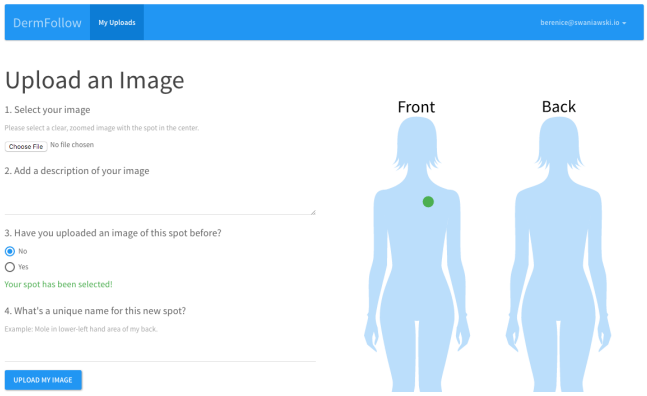

Figure 1: Patient Form

Figure 1 shows the patient interface for uploading an image of a skin lesion. Images are grouped by *spots*, which are unique skin lesions. A user can add a new image of a previously identified spot by choosing the existing spot, or interactively select a point on a 2D body map that identifies the new lesion. Once uploaded, the provider can see all of the patient's lesions on the body map, allowing easy interaction with the data (Figure 2).

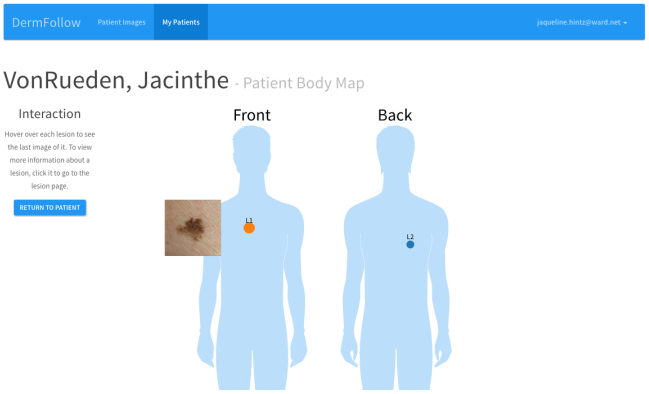

Figure 2: Provider View of Patient Body Map

Figure 3 displays the interface as seen by a dermatologist for a patient. The physician can easily see all the patient's images and their computed risk scores, as well as a global risk score for the patient.

We also provide an interface for the dermatologist (Figure 4) to view all his or her patients' recent uploads. If an upload is flagged as high-risk (malignant) by our CNN model, we display an asterisk next to the patient name above the

Figure 3: Provider View of Patient

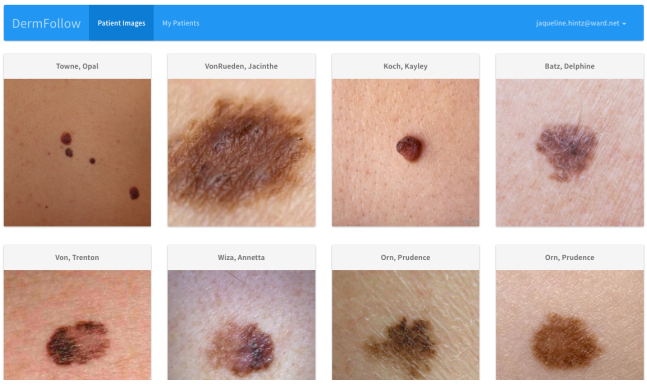

Figure 4: Provider View of Patient Uploads

image. We received substantial interest in this feature from the dermatologist community.

4.2.2 *Neural Network Architecture*

For classification, we use an ensemble of neural networks, including the VGG-16 [17], Inception [19] and ResNet [8] architectures. The VGG-16 and Inception models were fine-tuned for the task by lobotomizing the output layers of the network, adding an appropriately sized linear output layer and softmax layer, and retraining. The ResNet was trained de novo on the entirety of the publicly available International Skin Imaging Collaboration (ISIC) dermatological image dataset [9] over the course of two days. The 3,387 high-resolution images from this dataset were augmented 50 fold by a series of scale transformations, blurs, and rotations, that were precomputed for resource efficiency. The resulting 169,350 images were then used to train and cross-validate the models.

The final probability score of an uploaded image is the weighted average of scores from individual models, where the weight assigned to each model is equal to its testing accuracy. We obtained a binary classification (benign/malignant) accuracy of approximately 89.1 $\pm$ 1.8% from 10-fold cross-validation on our ensemble model.

4.2.3 *Risk Score*

For each patient, a risk score is computed which is based on a cutaneous lifetime melanoma risk algorithm from [4]:

$$risk\_score = \alpha \times image\_score + (1 - \alpha) \times info\_score \quad (1)$$

The *info_score* is computed from parameters such as skin type, hair color, and medical history. This contributes a relatively small proportion of the overall risk score. The *image_score* is a score computed by the neural network model. A risk score close to one indicates a high risk for developing skin cancer, whereas a score close to zero indicates relatively small risk.

Each new patient upload updates the *image_score* for the patient as a weighted average of the current *image_score*, and the probability score (determined from the neural network) of the new image.

It is computed as follows:

$$\begin{cases} (1 - 3\beta) \times score_{curr} + 3\beta \times score_{new\_image} & score > 0.5 \\ (1 - \beta) \times score_{curr} + \beta \times score_{new\_image} & score \leq 0.5 \end{cases}$$

The first update equation assigns a higher weight to the new image and decays the cumulative score rapidly. This ensures that if the new image is a high-risk image, the cumulative score is reflective of that change. Here, α and β are constants. We used a value of 0.8 and 0.25 for α and β respectively.

5. EXPERIMENT AND EVALUATION

5.1 Method

The study adheres to all guidelines for human subjects research. Patients were asked to fill out a pre-experiment questionnaire to obtain basic details such as age, gender, and medical history, as well as to consent to the use of their anonymized images for the study. This was done before providing them an account for DermFollow, to have a baseline understanding of their skin cancer risk and satisfaction with their dermatologist.

Following this, patients were briefed regarding the application and its features and were given a patient account. Patients uploaded multiple lesion images over the span of a week, after which they were asked to complete a post-experiment questionnaire regarding usability, integration into their daily activities, and value provided from using the application.

Clinician study participants were also given a one-week window to test the application, during which they reviewed patient image analyses and interacted with the visualizations. For clinical review of the application, we partnered with dermatologists Justin Ko, MD, MBA, of Stanford University Medical Center, and Laura Ferris, MD, PhD, from the University of Pittsburgh Medical Center.

5.2 Metrics

Post-experiment patient and provider surveys each contain approximately 20 Likert-scale questions, in which the patient or provider responds to a statement on a scale of one to five, with one being strongly disagree and five being strongly agree.

From this data, we will compute the mean, standard deviation, median, interquartile range, and p-value. We additionally ask three questions allowing long-form responses in which the patient and provider are asked to give their comments on the value provided from using the application.

We group questions into three key categories:

1. Usability/Performance of the Application
2. Workflow Integration
3. Value Provided
4. Applicability to Clinical Practice

6. PLAN OF ACTIVITIES

Going forward, Stefano is improving the model. Matt Cimino is running user studies. Matt May is launching the web application and assisting with user studies. Thanh is fine-tuning the risk score algorithms. Apurv is doing report generation and data analysis of user study results. Note: All team members have contributed equally to this work.

7. CONCLUSIONS AND DISCUSSION

(under construction)

8. APPENDIX

Many thanks to our wonderful physician-advisers, Justin Ko of Stanford and Ben Stoff of Emory Healthcare.

Shown in Figure 5 is our previous, and revised plan of activities.

Plan of Activities - Old

Person	Task
Matt. C	User Studies
Thanh	User Interface
Stefano	Model
Matt M.	Application
Apurv	Report/Analysis

Plan of Activities - Revised

Person	Task
Matt. C	Administration of User Studies
Thanh	Risk Scoring Algorithms
Stefano	Model Fine-Tuning, Validation
Matt M.	Web Application Launch, User Study Management
Apurv	Final Report Development / User Study Statistical Analysis

Figure 5: Plan of Activities

9. REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[2] R. Bostock, E. Claridge, A. Harget, and P. Hall. Towards a neural network based system for skin cancer diagnosis. In *Artificial Neural Networks, 1993., Third International Conference on*, pages 215–219. IET, 1993.

[3] C.-L. Chang and C.-H. Chen. Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, 36(2):4035–4041, 2009.

[4] J. R. Davies, Y.-m. Chang, D. T. Bishop, B. K. Armstrong, V. Bataille, W. Bergman, M. Berwick, P. M. Bracci, J. M. Elwood, M. S. Ernstoff, et al.

Development and validation of a melanoma risk score based on pooled data from 16 case–control studies. *Cancer Epidemiology Biomarkers & Prevention*, 24(5):817–824, 2015.

[5] F. Ercal, A. Chawla, W. V. Stoecker, H.-C. Lee, and R. H. Moss. Neural network diagnosis of malignant melanoma from color images. *Biomedical Engineering, IEEE Transactions on*, 41(9):837–845, 1994.

[6] A. Esteva, B. Kuprel, and S. Thrun. Deep networks for early stage skin disease and skin cancer classification. Unpublished manuscript, 2015.

[7] H. Ganster, A. Pinz, R. Röhrer, E. Wildling, M. Binder, and H. Kittler. Automated melanoma recognition. *Medical Imaging, IEEE Transactions on*, 20(3):233–239, 2001.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[9] ISIC. International skin imaging collaboration archive. https://isic-archive.com/, 2016.

[10] M. Kreutz, M. Anschütz, S. Gehlen, T. Grünendick, and K. Hoffmann. *Bildverarbeitung für die Medizin 2001: Algorithmen — Systeme — Anwendungen*, chapter Automated Diagnosis of Skin Cancer Using Digital Image Processing and Mixture-of-Experts, pages 357–361. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

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

[12] H. Lee and Y. P. Chen. Image based computer aided diagnosis system for cancer detection. *Expert Syst. Appl.*, 42(12):5356–5365, 2015.

[13] D. S. Rigel, R. J. Friedman, A. W. Kopf, and D. Polsky. Abcde—an evolving concept in the early detection of melanoma. *Archives of dermatology*, 141(8):1032–1034, 2005.

[14] P. Rubegni, G. Cevenini, M. Burroni, R. Perotti, G. Dell'Eva, P. Sbano, C. Miracco, P. Luzi, P. Tosi, P. Barbini, et al. Automated diagnosis of pigmented skin lesions. *International Journal of Cancer*, 101(6):576–580, 2002.

[15] M. A. Sheha, M. S. Mabrouk, and A. Sharawy. Automatic detection of melanoma skin cancer using texture analysis. *International Journal of Computer Applications*, 42(20):22–26, 2012.

[16] S. Sigurdsson, P. A. Philipsen, L. K. Hansen, J. Larsen, M. Gniadecka, and H. C. Wulf. Detection of skin cancer by classification of raman spectra. *Biomedical Engineering, IEEE Transactions on*, 51(10):1784–1793, 2004.

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

[18] S. M. Strayer and P. Reynolds. Diagnosing skin malignancy: assessment of predictive clinical criteria and risk factors.(research findings that are changing clinical practice). *Journal of family practice*, 52(3):210–219, 2003.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR 2015*, 2015.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[21] L. Thomas, P. Tranchand, F. Berard, T. Secchi, C. Colin, and G. Moulin. Semiological value of abcde criteria in the diagnosis of cutaneous pigmented tumors. *Dermatology*, 197(1):11–17, 1998.

[22] E. Zagrouba and W. Barhoumi. A prelimary approach for the automated recognition of malignant melanoma. *Image Analysis and Stereology Journal*, 23(2):121–135, 2004.