PRECISION DIAGNOSIS OF MELANOMA AND OTHER SKIN LESIONS FROM DIGITAL IMAGES

Abhishek Bhattacharya [1], Albert Young [1,2], Andrew Wong [1,2], Simone Stalling [1], Maria Wei [3,4], Dexter Hadley [1,2,3]

[1] Institute for Computational Health Sciences, [2] Department of Pediatrics, [3]School of Medicine, [4]Department of Dermatology,
University of California, San Francisco, CA 94143

Abstract

Melanoma will affect an estimated 73,000 new cases this year and result in 9,000 deaths, yet precise diagnosis remains a serious problem. Without early detection and preventative care, melanoma can quickly spread to become fatal (Stage IV 5-year survival rate is 20-10%) from a once localized skin lesion (Stage IA 5-year survival rate is 97%). There is no biomarker for melanoma in clinical use, and the current diagnostic criteria for skin lesions remains subjective and imprecise. Accurate diagnosis of melanoma relies on a histopathologic gold standard; thus, aggressive excision of melanocytic skin lesions has been the mainstay of treatment. It is estimated that 36 biopsies are performed for every melanoma confirmed by pathology among excised lesions. There is significant morbidity in misdiagnosing melanoma such as progression of the disease for a false negative prediction vs the risks of unnecessary surgery for a false positive prediction. Every year, poor diagnostic precision adds an estimated $673 million in overall cost to manage the disease.

Currently, manual dermatoscopic imaging is the standard of care in selecting atypical skin lesions for biopsy, and at best it achieves 90% sensitivity but only 59% specificity when performed by an expert dermatologist. Many computer vision (CV) algorithms perform better than dermatologists in classifying skin lesions although not significantly so in clinical practice. Meanwhile, open source deep learning (DL) techniques in CV have been gaining dominance since 2012 for image classification, and today DL can outperform humans in classifying millions of digital images with less than 5% error rates. Moreover, DL algorithms are readily run on commoditized hardware and have a strong online community of developers supporting their rapid adoption. In this work, we performed a successful pilot study to show proof of concept to DL skin pathology from images.

However, DL algorithms must be trained on very large labelled datasets of images to achieve high accuracy. Here, we begin to assemble a large imageset of skin lesions from the UCSF and the San Francisco Veterans Affairs Medical Center (VAMC) dermatology clinics that are well characterized by their underlying pathology, on which to train DL algorithms. If trained on sufficient data, we hypothesize that our approach will significantly outperform general dermatologists in predicting skin lesion pathology. We posit that our work will allow for precision diagnosis of melanoma from widely available digital photography, which may optimize the management of the disease by decreasing unnecessary office visits and the significant morbidity and cost of melanoma misdiagnosis.

Background and Significance

Melanoma misdiagnosis is a significant public health problem

Over the last two decades the number of patients in the United States diagnosed with melanoma has steadily risen to make it the fifth most common cancer in the nation. This year alone, an estimated 73,000 new cases and 9,000 deaths are expected to occur due to the disease [1]. Although early stages are highly survivable (Stage IA 5-year survival rate is 97%), without early detection and preventative care, melanoma can quickly spread and become fatal (Stage IV 5-year survival rate is 20-10%) [2,3]. Poor diagnostic precision adds an estimated $673 million in overall cost to the management of the disease[4–6]

Melanoma pathophysiology and staging

Melanoma typically arises in pigment-producing cells known as melanocytes that have undergone adverse genetic mutation most frequently attributed to ultraviolet light (UV) radiation exposure [7–9]. Aside from UV exposure, some rare hereditary mutations in genes such as *CDKN2A*, *CDK4*, and *MC1R* can also be good indicators for patients with high risk of developing familial melanoma (patients that have families with a history

of melanoma) [10]. The progression of the disease is best characterized both clinically and histopathologically, and it can rapidly progress from stage 0 (melanoma *in situ*) to stage 4 (metastatic melanoma) [11] without proper diagnosis and management.

Accurate clinical diagnosis of melanoma can be challenging

The ABCDE method for visually assessing pigmented skin lesions for malignancy [12] outlines Asymmetry, Border irregularity, Color variegation, Diameter (>6mm), and Evolution as clinical features to follow [13]. However, diagnostic accuracy of melanoma by the unaided eye is disappointing [14,15]. The melanoma yield on biopsy of suspicious lesions is only 1 in 36 [16]. Dermoscopy [17] facilitates visualization of morphological features which are not discernible by examination with the naked eye [18], and it enables better diagnosis as compared to unaided eye [19–21] with an improvement in diagnostic sensitivity of 10–30% [22]. However, dermoscopy may actually lower the diagnostic accuracy in the hands of inexperienced dermatologists [23–26], since this method requires great deal of experience to differentiate skin lesions [27]. Experts achieve 90% sensitivity and 59% specificity, while this performance significantly worsens with inexperience and drops to 62%-63%for general practitioners [28,29]. Currently available computer vision (CV) algorithms perform only marginally better in practice with no significant improvement in diagnosis relative to a dermatologist [30]. Histopathology remains the gold standard for accurate melanoma diagnosis [31] although the rate of discordant readings between pathologists can be high: when 11 expert pathologists reviewed 37 'classic' melanocytic lesions there was total agreement in only 30% of cases; other studies report up to a 50% discordance rate among pathologists [32–36]. Thus the diagnostic accuracy of melanoma remains problematic independent of the method used for diagnosis.

Deep learning is emerging because of Big Data

Deep learning (DL) emerged from the traditional neural network paradigm of artificial intelligence that was developed in the 1980s to computationally model neuronal activity in the brain. An artificial neuron is modeled to fire around an activation threshold (or bias) and differentially weighted inputs. However, to interpret complex signals and patterns requires sophisticated models of computational neurons that are chained together to propagate signals much like the visual system in brain interprets light signals with successive cognitive interpretation (retina, V1, V2, etc.) in order to classify objects. Today, the most useful neural network models are composed of thousands of multi-layered artificial neurons that are parameterized by exponentially more biases and weights that require massive datasets to estimate. However, once these networks are trained on sufficiently large high quality labeled datasets, they generally outperform other machine learning methods. The computationally intensive process of accurately estimating their parameters by training on massive datasets constitutes the paradigm of DL. Furthermore, the exponential growth in computational power and the recent emergence of GPU computation, together with the abundance of large data sets to train on makes DL application more practical now than ever before.

Deep Learning facilitates the most accurate image classification

Big data, cheap computation, and better algorithms are making breakthroughs in artificial intelligence such as deep learning (DL) more possible now than ever before [37]. Today, DL is being applied to a variety of tasks with extraordinary results [38–53], but the most remarkable progress has been made in the field of computer vision (CV). ImageNet holds an annual Large Scale Visual Recognition Challenge (ILSVRC) competition [54,55] for teams to classify 1.2 million images of objects into 1,000 categories (). In 2010, all teams used traditional CV algorithms with accuracy rates <71.8%. Only Incremental progress was

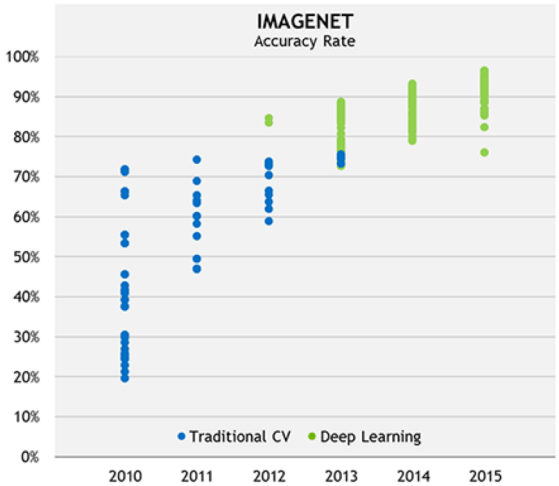

Figure 1: Deep Learning far outperforms traditional computer vision in image recognition. The figure shows the relative performance improvement in error rates on ImageNet's Large Scale Visual Recognition Challenge (ILSVRC) to classify one million images into one thousand categories

made until 2012 when Alex Krizhevsky *et al.* submitted the AlexNet [56] DL approach with a 83.6% accuracy rate that far outperformed the competition. By ILSVRC 2013, all other participants embraced DL, and Google won ILSVRC 2014 with its original Inception model architecture [57]. Since then, Microsoft first outperformed humans with > 95% accuracy rates classifying the ImageNet dataset [58] and Google has subsequently leap-frogged to lead image classification performance with its latest open Inception v3 model [59] that achieves > 96.5% accuracy rates. This proposal is innovative because it leverages the impressive performance of general state-of-the-art deep learning models to improve the current standards of digital mammography screening.

Open source DL frameworks are becoming popular

This project will utilize popular open source deep learning frameworks such as Caffe [60], Theano [61,62], and Torch [63]. All of these frameworks have contributed to numerous publications and are implemented everywhere from academia to industry. Many of the state-of-the-art algorithms that have won computer vision competitions in the past are published in public repositories tied to each of these frameworks. For example, Caffe, the deep learning framework developed out of the Berkeley Vision Lab, has their own "Model Zoo" where researchers and community members can publish and share pre-trained models. The repository not only includes the first model used to win ImageNet in 2012 [58], but also the latest networks Google uses for their own large scale image classification tasks [57]. The active community of developers that are supporting these open tools will be a valuable resource to fine-tuning pre-existing models as well as to build the novel DL architecture to accurately diagnose skin lesion pathology as we propose here.

Figure 2: Example Malignant melanoma photo from the DermNet Skin Atlas

Results

We performed a successful pilot study to show proof of concept of DL skin pathology from publically available images.

Publically available digital images of skin lesions

Google crawls the Internet to catalog every available image online, and we searched Google Images for "melanocytic skin lesion". The majority of images of melanocytic skin lesion came from DermNet Skin Atlas [64], the largest independent photo dermatology source dedicated to online medical education. However only low resolution watermarked lesions (Figure 2) were freely available to download while high resolution images were available at $50 / image. We scraped freely available low quality images from DermNet and labeled them with DermNet assigned diagnoses. In all we identified 275 images labeled by DermNet comprising 170 atypical lesions (42 atypical nevi + 128 malignant melanoma) and 115 benign lesions (25 halo nevi + 80 melanocytic nevi).

Figure 3: TSNE plot of 275 images

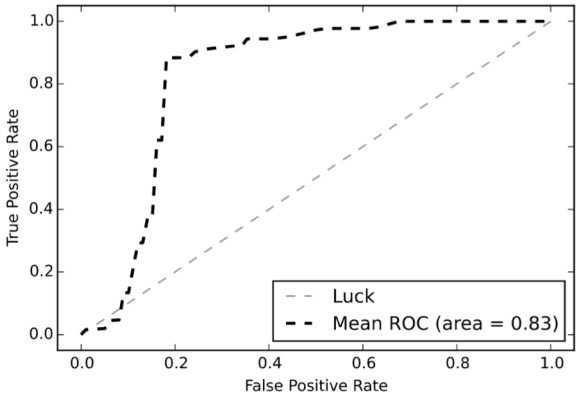

Figure 4: Receiver operating curve for predicting skin pathology from 275 images with 10 fold cross validation.

Deep Learning features of digital images

We utilized a type of algorithm called transfer learning to train a DL algorithm to classify skin lesions as typical or atypical. Transfer learning is applicable to our limited dataset on standard personal computer hardware because it assumes

a DL reference model for feature selection, with subsequent classification of this imageset by traditional machine learning algorithms. We utilized the BVLC AlexNet model from Caffe Model Zoo to reduce the pixels of each digital image to 4,096 features. We used the Python programming language with scikit-learn open libraries to project the 4,096 features learned from the DL algorithm into 2 dimensions with the TSNE method in order to visualize the relationship among 275 images (Figure 3). We found that even with the low quality watermarked DermNet images, were were able to identify structure in the relatedness of images. Specifically, nevi clustered together as did images with hair, as well as different subsets of melanoma. Finally, to assess the predictive power of the features extracted by DL, we used a support vector machine (SVM) classifier trained on the 275 DermNet images with parameter optimization by cross validation. We found the area under the curve (AUC) of 0.83 after 10-fold cross validation of the SVM (Figure 4), and an AUC of of 0.80 to 0.90 represents a "good" measure of accuracy by most standards. We expect that 0.83 is the lower bound on the accuracy of our approach as we will increase our predictive power by training more sophisticated models with much more data.

Conclusion

At this time, computers cannot replace an experienced clinician's intuition. However, with proficient training on sufficient high-quality data, CV algorithms will eventually match, if not exceed, clinical diagnostic accuracy of dermatologists. Applying DL to a large well-characterized prospectively collected clinical imagesets may indeed yield new diagnostic tools to more accurately diagnose skin cancer. Precision diagnostics of melanoma may serve as a first step in significantly reducing the mortality rate and improving the overall management of the disease.

Literature Cited

1. American Cancer Society. Cancer Facts & Figures 2015. (2015). doi:10.3322/caac.21254
2. Dna, A. Melanoma Skin Cancer What is cancer ? *Am. Cancer Soc.* (2013).
3. Balch, C. M. *et al.* Final version of 2009 AJCC melanoma staging and classification. *J. Clin. Oncol.* **27,** 6199–6206 (2009).
4. Gordon, L. *et al.* Diagnosis and management costs of suspicious skin lesions from a population-based melanoma screening programme. *J. Med. Screen.* **14,** 98–102 (2007).
5. Tsao, H., Rogers, G. S. & Sober, A. J. An estimate of the annual direct cost of treating cutaneous melanoma. *J. Am. Acad. Dermatol.* **38,** 669–680 (1998).
6. Inflation Calculator: Bureau of Labor Statistics. at <http://www.bls.gov/data/inflation_calculator.htm>
7. Whiteman, D. & Green, A. The pathogenesis of melanoma induced by ultraviolet radiation. *The New England journal of medicine* **341,** 766–767 (1999).
8. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* (2012). doi:10.1038/nature11071
9. Mouret, S. *et al.* Cyclobutane pyrimidine dimers are predominant DNA lesions in whole human skin exposed to UVA radiation. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 13765–70 (2006).
10. Mize, D. E., Bishop, M., Resse, E. & Sluzevich, J. Familial Atypical Multiple Mole Melanoma Syndrome. (2009).
11. Leong, S. P. L. *et al.* Progression of cutaneous melanoma: implications for treatment. *Clin. Exp. Metastasis* **29,** 775–96 (2012).
12. Rigel, D. S., Friedman, R. J., Kopf, A. W. & Polsky, D. ABCDE--an evolving concept in the early detection of melanoma. *Arch. Dermatol.* **141,** 1032–4 (2005).
13. Wurm, E. M. & Peter Soyer, H. Scanning for melanoma. *Aust. Prescr.* **33,** 150–155 (2010).

14. Morton, C. A. & Mackie, R. M. Clinical accuracy of the diagnosis of cutaneous malignant melanoma. *Br. J. Dermatol.* **138,** 283–287 (1998).

15. Lindelöf, B. & Hedblad, M. A. Accuracy in the clinical diagnosis and pattern of malignant melanoma at a dermatological clinic. *J. Dermatol.* **21,** 461–464 (1994).

16. Bataille, V., Sasieni, P., Curley, R. K., Cook, M. G. & Marsden, R. A. Melanoma yield, number of biopsies and missed melanomas in a British teaching hospital pigmented lesion clinic: A 9-year retrospective study. *Br. J. Dermatol.* **140,** 243–248 (1999).

17. Argenziano, G. & Soyer, H. P. Dermoscopy of pigmented skin lesions--a valuable tool for early diagnosis of melanoma. *Lancet Oncol.* **2,** 443–449 (2001).

18. Masood, A. & Al-Jumaily, A. A. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *Int. J. Biomed. Imaging* **2013,** 323268 (2013).

19. Carli, P. *et al.* Addition of dermoscopy to conventional naked-eye examination in melanoma screening: A randomized study. *J. Am. Acad. Dermatol.* **50,** 683–689 (2004).

20. Carli, P. *et al.* Improvement of malignant/benign ratio in excised melanocytic lesions in the 'dermoscopy era': a retrospective study 1997-2001. *Br. J. Dermatol.* **150,** 687–92 (2004).

21. Vestergaard, M. E., Macaskill, P., Holt, P. E. & Menzies, S. W. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *Br. J. Dermatol.* **159,** 669–76 (2008).

22. Mayer, J. Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. *Med J Aust* **167,** 206–10. (1997).

23. Binder, M. *et al. Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. Archives of dermatology* **131,** (1995).

24. Piccolo, D. *et al.* Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: A comparative study. *Br. J. Dermatol.* **147,** 481–486 (2002).

25. Braun, R. P., Rabinovitz, H. S., Oliviero, M., Kopf, A. W. & Saurat, J. H. Dermoscopy of pigmented skin lesions. *Journal of the American Academy of Dermatology* **52,** 109–121 (2005).

26. Kittler, H., Pehamberger, H., Wolff, K. & Binder, M. Diagnostic accuracy of dermoscopy. *Lancet Oncology* **3,** 159–165 (2002).

27. Whited, J. D. Does This Patient Have a Mole or a Melanoma? *JAMA: The Journal of the American Medical Association* **279,** 696–701 (1998).

28. Menzies, S. W. *et al.* The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. *Arch. Dermatol.* **141,** 1388–1396 (2005).

29. Burroni, M. *et al.* Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study. *Clin. Cancer Res.* **10,** 1881–1886 (2004).

30. Rajpara, S. M., Botello, A. P., Townend, J. & Ormerod, A. D. Systematic review of dermoscopy and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. *Br. J. Dermatol.* **161,** 591–604 (2009).

31. Smoller, B. R. Histologic criteria for diagnosing primary cutaneous malignant melanoma. *Mod. Pathol.* **19 Suppl 2,** S34–S40 (2006).

32. Ferrara, G. *et al.* The influence of clinical information in the histopathologic diagnosis of melanocytic skin neoplasms. *PLoS One* **4,** (2009).

33. Lodha, S., Saggar, S., Celebi, J. T. & Silvers, D. N. Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting. *J. Cutan. Pathol.* **35,** 349–352 (2008).

34. Farmer, E. R., Gonin, R. & Hanna, M. P. Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. *Hum. Pathol.* **27,** 528–531 (1996).

35. Corona, R. *et al.* Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions. *J. Clin. Oncol.* **14,** 1218–23 (1996).

36. Ackerman, A. B. Discordance among expert pathologists in diagnosis of melanocytic neoplasms. *Hum. Pathol.* **27,** 1115–6 (1996).

37. The Three Breakthroughs That Have Finally Unleashed AI on the World | WIRED. at <http://www.wired.com/2014/10/future-of-artificial-intelligence/>

38. Ba, J. L., Mnih, V. & Kavukcuoglu, K. Multiple Object Recognition With Visual Attention. *Iclr* 1–10 (2015).

39. Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P. J. & Gonzalez-Rodriguez, J. Frame-by-frame language identification in short utterances using deep neural networks. *Neural Networks* **64,** 49–58 (2015).

40. Nair, A. *et al.* Massively Parallel Methods for Deep Reinforcement Learning. *arXiv:1507.04296* 14 (2015).

41. Angelova, A., Krizhevsky, A. & Vanhoucke, V. Pedestrian detection with a Large-Field-Of-View deep network. in *Proceedings - IEEE International Conference on Robotics and Automation* **2015–June,** 704–711 (2015).

42. Heigold, G. *et al.* Multilingual Acoustic Models using Distributed Deep Neural Networks. in *Icassp* 8619–8623 (2013). doi:10.1109/ICASSP.2013.6639348

43. Hinton, G. *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Process. Mag.* 82–97 (2012). doi:10.1109/MSP.2012.2205597

44. Zeiler, M. D. *et al.* On rectified linear units for speech processing. in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 3517–3521 (2013). doi:10.1109/ICASSP.2013.6638312

45. Karpathy, A. *et al.* Large-scale video classification with convolutional neural networks. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1725–1732 (2014). doi:10.1109/CVPR.2014.223

46. Szegedy, C. *et al.* Going Deeper with Convolutions. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **07–12–June,** 1–9 (2014).

47. Frome, A., Corrado, G. & Shlens, J. Devise: A deep visual-semantic embedding model. *Adv. Neural …* 1–11 (2013).

48. Vinyals, O. *et al.* Grammar as a Foreign Language. *arXiv* 1–10 (2014). doi:10.1146/annurev.neuro.26.041002.131047

49. Mikolov, T., Corrado, G., Chen, K. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *Proc. Int. Conf. Learn. Represent. (ICLR 2013)* 1–12 (2013). doi:10.1162/153244303322533223

50. Le, Q. V *et al.* Building high-level features using large scale unsupervised learning. *Int. Conf. Mach. Learn.* 38115 (2011). doi:10.1109/MSP.2011.940881

51. Ramsundar, B. *et al.* Massively Multitask Networks for Drug Discovery. (2015).

52. Lusci, A., Pollastri, G. & Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* **53,** 1563–75 (2013).

53. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).

54. Deng, J. D. J. *et al.* ImageNet: A large-scale hierarchical image database. *2009 IEEE Conf. Comput. Vis. Pattern Recognit.* 2–9 (2009). doi:10.1109/CVPR.2009.5206848

55. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115,** 211–252 (2015).

56. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* 1097–1105 (2012).

57. Szegedy, C. *et al.* Going Deeper with Convolutions. (2014).

58. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.

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

60. Jia, Y. *et al.* Caffe: Convolutional Architecture for Fast Feature Embedding. (2014).

61. Bergstra, J. *et al.* Theano: A CPU and GPU Math Compiler in Python. in *Proceedings of the 9th Python in Science Conference* 3–10 (2010).

62. Bastien, F. *et al.* Theano: new features and speed improvements. (2012).

63. Torch | Scientific computing for LuaJIT. at <http://torch.ch/>

64. Dermatology Education | Just another WordPress site. at <http://www.dermnet.com/>