Research Article

Classification of Melanocytic Lesions in Selected and Whole-Slide Images via Convolutional Neural Networks

Steven N. Hart[1], William Flotte[2], Andrew P. Norgan[2], Kabeer K. Shah[2], Zachary R. Buchan[2], Taofic Mounajjed[2], Thomas J. Flotte[2]

[1]Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo College of Medicine, Rochester, Minnesota, USA,
[2]Department of Laboratory Medicine and Pathology, Mayo College of Medicine, Rochester, Minnesota, USA

Received: 18 May 2018 Accepted: 27 November 2018 Published: 20 February 2019

Abstract

Whole-slide images (WSIs) are a rich new source of biomedical imaging data. The use of automated systems to classify and segment WSIs has recently come to forefront of the pathology research community. While digital slides have obvious educational and clinical uses, their most exciting potential lies in the application of quantitative computational tools to automate search tasks, assist in classic diagnostic classification tasks, and improve prognosis and theranostics. An essential step in enabling these advancements is to apply advances in machine learning and artificial intelligence from other fields to previously inaccessible pathology datasets, thereby enabling the application of new technologies to solve persistent diagnostic challenges in pathology. Here, we applied convolutional neural networks to differentiate between two forms of melanocytic lesions (Spitz and conventional). Classification accuracy at the patch level was 99.0%–2% when applied to WSI. Importantly, when the model was trained without careful image curation by a pathologist, the training took significantly longer and had lower overall performance. These results highlight the utility of augmented human intelligence in digital pathology applications, and the critical role pathologists will play in the evolution of computational pathology algorithms.

Keywords: Bioinformatics, deep learning, dermatology, image analysis

Introduction

Melanocytic nevi are mostly benign and common, but certain forms of nevi can be difficult to classify; however, accurate classification of nevi is important in feature evaluation for distinguishing nevi from melanoma. The architecture and cytomorphology of different types of nevi vary significantly and their overlapping characteristics further confound the accurate diagnosis of malignancy. Features of malignant lesions are also found in benign nevi,[1] which makes diagnosis difficult. Depending on the criteria, accurate diagnoses range from 71% to 82%,[2] leading to 17.6% false diagnoses of melanoma.[3]

Recently, whole-slide image (WSI) scanners have made it possible to fully digitize pathology slides. In addition to enabling long-term slide preservation and facilitating slide sharing for collaboration or second opinions, digitization of pathology slides allows for the development and utilization artificial intelligence (AI)-driven diagnostic tools. During microscopic examination, a pathologist uses salient clinical information, pattern matching, and feature recognition (shape, color, structure, etc.) to render a diagnosis. For example, to diagnose melanoma, relevant features may include asymmetry, poor circumscription, predominance of single melanocytes, mitoses, necrosis, and other features. The major objective of this study was to develop a convolutional neural network (CNN) capable of distinguishing between conventional and Spitz nevi. A classification challenge exists in the diagnosis of a subset of melanocytic nevi as conventional or Spitz-type; a difficult but clinically important task. To accomplish this, curated image patches of conventional nevi, Spitz nevi, or nonnevus skin tissue (other) were manually extracted from WSIs by a board-certified dermatopathologist. The curated patches were used to train a CNN for the classification task.

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/jpi.jpi_32_18

Address for correspondence: Dr. Steven N. Hart,
Department of Health Sciences Research, Division of Biomedical Statistics
and Informatics, Mayo College of Medicine, Rochester, Minnesota, USA.
E-mail: hart.steven@mayo.edu

For reprints contact: reprints@medknow.com

How to cite this article: Hart SN, Flotte W, Norgan AP, Shah KK, Buchan ZR, Mounajjed T, *et al*. Classification of melanocytic lesions in selected and whole-slide images via convolutional neural networks. J Pathol Inform 2019;10:5.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2019/10/1/5/252613

Methods

We investigated the utility of a CNN to assist in the classification of selected melanocytic lesions as Spitz or conventional. Histologic sections of pigmented lesions were reviewed by two board-certified dermatopathologists and only cases where there was concurrence of diagnoses of conventional and Spitz nevi were utilized. Slides were digitized using an Aperio AT Turbo scanner from Leica Biosystems, with ×40 power. Large sections of representative tissue were curated by an expert dermatopathologist from 300 hematoxylin and eosin (H and E) slides each containing conventional ($n = 150$) or Spitz nevi [$n = 150$, Figure 1]. Slides were digitized using an Aperio AT Turbo scanner from Leica Biosystems, with ×40 power. The scans from 100 H and E slides (50 conventional and 50 Spitz nevi) were used for the training and validation set. Smaller variant image patches 299 × 299 pixels (px) were then derived for conventional nevi ($n = 15{,}868$), Spitz nevi ($n = 21{,}468$), and other nonnevus skin features ($n = 38{,}374$). From these patches, 30% were used exclusively for validation experiments. These image sets were then used to train and validate the deep CNN (Inception V3[4]) using the TensorFlow framework (version: 1.5.0).[5] Models were trained using pretrained weights (available from the TensorFlow website) or entirely from scratch. Using pretrained weights decrease the time to convergence since it reuses the weights that identify sample agnostic image characteristics such as edges and curves. Training from scratch means that the weights are initially randomized and then adjusted throughout the training process to converge. This process typically yields higher accuracy but requires more data and compute time to relearn basic features in addition to sample-dependent features (e.g., nuclei, cells, tissue compartments). In both cases, we used the following hyperparameters: RMSprop optimizer, batch size of 32, learning rate of 0.01, and training for 250k steps. At 250k steps, the model observed each image approximately 150 times (epochs).

A second experiment was also performed on noncurated image patches representing the entire slide. In this experiment, tissue segments were automatically extracted from the WSI without pathologist input. Successive nonoverlapping 299 × 299 px tiles representing the entire WSI were evaluated for tissue content by converting the red, green, and blue values to gray scale and applying a mean intensity cut-off of >210. Any 299 × 299 px region with sufficient gray scale intensity was considered to possibly contain tissue and was extracted and analyzed. Regions with insufficient gray scale intensity were not considered for the analysis and treated as missing data. Since no human selection occurred, only two prediction classes were available: Spitz and conventional, with $n = 611{,}485$ and $n = 612{,}523$ image patches, respectively, from the 100 training slides. To effectively compare the results to the curated patch-level classifications, training was performed for 3.6 M steps (~135 epochs).

Testing was performed using 200 WSI not used during training or validation. Accuracy was measured at the patch level (from the validation patches) and at the WSI level. WSI were classified as either conventional or Spitz by calculating a prediction for all nonoverlapping 299 × 299 px regions with sufficient tissue. Classifications where the classification probability (i.e., logit) was at least 10% higher than the next likely class were used as votes, with the classification label for the entire slide assigned by simple majority (Spitz or conventional, [other was ignored]). Accuracy for the WSI-label predictions was then assessed for binary classification accuracy using the Caret package (version: 6.0–71)[6] in R (version: 3.2.3).[7] The gold standard for the correct classification was the diagnosis made by the dermatopathologist.

All codes used for these data are publicly available on GitHub.[8] This work was conducted under approval from the Institutional Review Board at Mayo Clinic.

Results

Training using the curated image patches took approximately 50 h to complete 250k iterations with 4 GeForce GTX 1080 GPUs. Training accuracy for curated patches reached maximum accuracy (100%) at around epoch 13, whereas the pretrained model only began to converge around epoch 100 [Figure 2]. Training accuracy for the noncurated patches converged around epoch 50. The validation accuracy, however, revealed stark differences in the generalizability of the models.

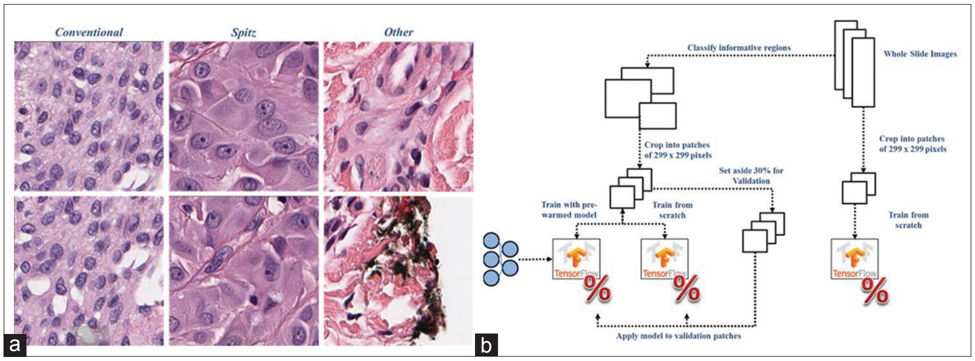
Conventional
Spitz
Other
a
Classify informative regions
Whole Slide Images
Crop into patches of 299 x 299 pixels
Set aside 30% for Validation
Crop into patches of 299 x 299 pixels
Train with pre-warmed model
Train from scratch
Train from scratch
TensorFlow
%
TensorFlow
%
TensorFlow
%
Apply model to validation patches
b

Figure 1: Experimental design. (a) Representative examples of image classes. (b) Sample image selection and modeling. Note the “other” class was only available for the curated informative regions

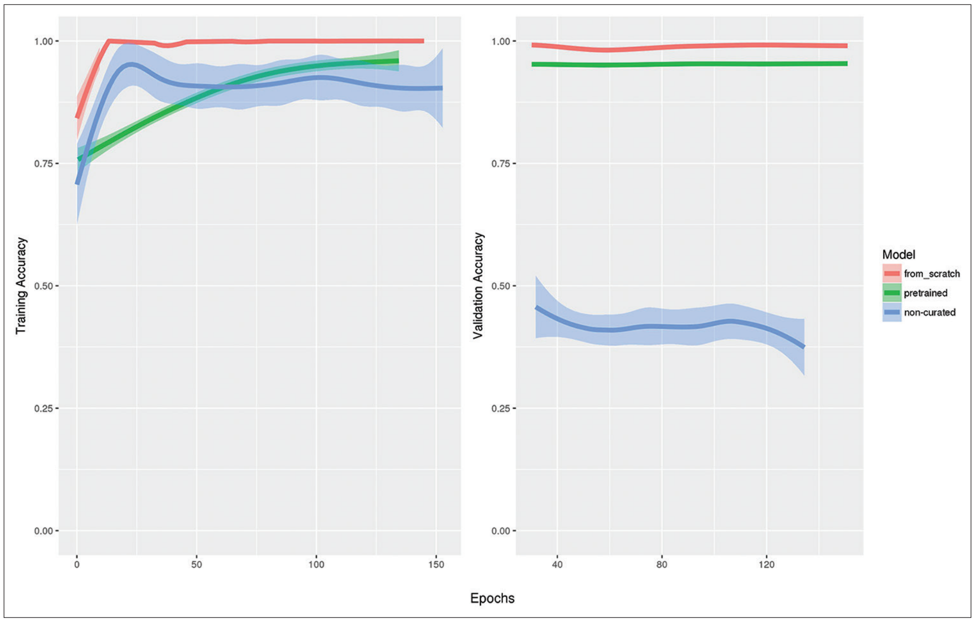

Figure 2: TTraining and Validation Accuracy. (Left) Training accuracy for each cohort of images and models. The shaded area is the margin of error. (Right) Accuracy of predictions on the validation images

Both the *de novo* and pretrained networks had high validation accuracy (99.0% and 95.4%, respectively), but the noncurated patches were unable to learn transferable features with a final validation accuracy of only 52.3%.

A single classification was applied to an entire slide denoting whether or not it contained a Spitz or conventional nevus. For each patch in a given WSI, a prediction was made as to whether that patch was of type "Spitz," "conventional," or "other." Then, the number of patches that were predicted as Spitz or conventional was tallied, and an overall slide prediction was based on whichever category was more abundant. That WSI-level prediction is then compared to the true label of the slide to determine accuracy. The classification accuracy of the 200 whole slides not seen by the training algorithm was 92.0%. Sensitivity was 85% with a specificity of 99%. On a per class basis, 99 of 100 conventional nevi were classified correctly (99%), compared to only 85% for Spitz nevi. Of the 16 misclassified WSI, 94% were due to Spitz-type lesions being classified as conventional. When further exploring the false-positive calls, a strong edge effect was observed around the decision boundary [Figure 3], meaning that the incorrect calls were primarily driven by small differences in the expected versus observed classes. Examples of correctly and incorrectly predicted WSI are shown in Figure 4.

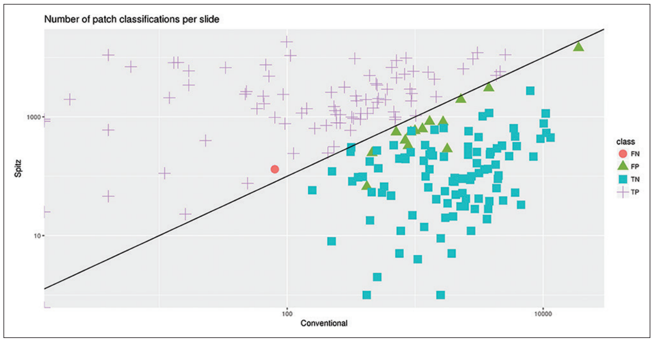

Figure 3: Experimental design. Count of patch predictions from the whole-slide image. For each whole-slide image, the total number of predictions for Spitz and conventional was aggregated. Squares and crosses signify correct classifications. Circles and triangles are misclassified whole-slide image. Notice the majority of misclassified images reside near the decision boundary (solid line)

Conclusions

This work highlights an important lesson when developing algorithms for use by pathologists; involve the pathologist in the design of the assay. The manual curation, though tedious for the clinician, proved to be a valuable contribution to optimizing model performance. By preselecting representative examples of Spitz and conventional nevi, along with providing examples for nondiagnostic areas such as hair follicles, sweat glands, and tissue artifacts, the model was able to learn faster and has an overall higher accuracy on the training and validation sets with fewer examples. The number of images used from the curated images was ×16 less than the noncurated approach but was more focused on learning the salient features for discrimination in less time, taking only 50 h to train versus 800.

Given a small image patch, the algorithm will correctly predict the correct classification 99% of the time. However, there are several important caveats. At present, the classifier does not achieve high accuracy with undirected evaluation of all image patches extracted from WSIs to be reliable for clinical use. Our data show that the accuracy of a single call for a WSI is 92% accurate. This is a major limitation since this would be the expected workflow in clinical practice.

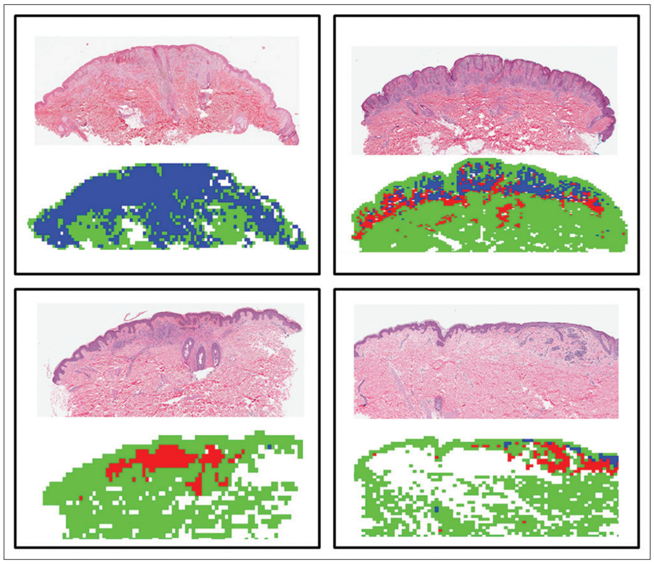

Figure 4: Example classification of the whole-slide image. Each of these images shows an example of correct (left) or incorrect (right) classification for Spitz (top) and conventional (bottom) nevi types. In the heatmaps adjacent to each image, each pixel is colored to represent the prediction for a particular region. Blue indicates a patch-level classification for "Spitz," red for "conventional," and green for "other"

These errors are predominantly derived around a decision boundary between the number of patches counted as either Spitz or conventional. More sophisticated methods will be needed to improve classification accuracy at the whole-slide level. Alternatively, more work could be done to improve the patch-level accuracy (currently at 99%), which would decrease the number of false calls in a WSI. Given that each WSI generates about 15,000 image patches, a 1% error rate would result in 150 false-positive calls, which on its face does not seem alarming. However, ~75% of those fall into the "other," noninformative classification, so the influence of even a few incorrect assertions can have moderate influence on the final classification. Additional work on refining that initial classification or on developing a secondary machine learning framework for results interpretation is necessary to decrease the error rate of diagnostic classification of Spitz versus conventional Nevi. These data provide strong evidence for the potential utility of AI to enhance diagnosis in digital pathology.

Acknowledgments

The authors would like to thank the Department of Laboratory Medicine and Pathology at Mayo Clinic for funding this work.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Urso C, Rongioletti F, Innocenzi D, Batolo D, Chimenti S, Fanti PL, *et al.* Histological features used in the diagnosis of melanoma are frequently found in benign melanocytic naevi. J Clin Pathol 2005;58:409-12.
2. Annessi G, Bono R, Sampogna F, Faraggiana T, Abeni D. Sensitivity, specificity, and diagnostic accuracy of three dermoscopic algorithmic methods in the diagnosis of doubtful melanocytic lesions: The importance of light brown structureless areas in differentiating atypical melanocytic nevi from thin melanomas. J Am Acad Dermatol 2007;56:759-67.
3. Brochez L, Verhaeghe E, Grosshans E, Haneke E, Piérard G, Ruiter D, *et al.* Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions. J Pathol 2002;196:459-66.
4. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016.
5. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems; 2016. Available from: http://www.arxiv.org/abs/1603.04467. [Last accessed on 2018 Aug 01].
6. CRAN – Package Caret. Available from: https://www.CRAN.R-project.org/package=caret. [Last accessed on 2018 Mar 28].
7. R: The R Project for Statistical Computing. Available from: https://www.R-project.org/. [Last accessed 2018 on Mar 28].
8. Steven-N-Hart. Steven-N-Hart/WSI-Classification. In: GitHub. Available from: https://www.github.com/Steven-N-Hart/WSI-Classification. [Last accessed on 2018 Mar 27].