Deep Vision: Learning to Identify Renal Disease With Neural Networks

Check for updates

Nishanth P. Pavinkurve[1], Karthik Natarajan[1] and Adler J. Perotte[1]

[1]Department of Biomedical Informatics, Columbia University, New York, New York, USA

See Clinical Research on Page 955

Machine learning refers to building and using mathematical models that can effectively perform a specific task without using explicit instructions, relying on pattern recognition from existing examples instead. With advancements in computational capacity and efficient models in recent years, patterns that were not easily noticed by most humans are exploited by machines, resulting in their performances on par, if not better, than humans in several tasks. In health care specifically, machine learning for computer vision has been successfully used in detecting diabetic retinopathy from photographs,[1] analyzing biopsy slides,[2] magnetic resonance imaging,[3] computed tomography scans,[4] and radiographs.[5] The authors have previously worked with machine learning approaches for analyzing biopsy images to predict the stage of chronic kidney disease, among other outcomes, and have continued to explore machine learning approaches on biopsies in this current work.[6,7] In recent years, because of the enormous number of laboratory reports and image data being generated, and consequently the burnout clinicians are facing, machine learning offers a potential solution to keep up with the volume; however, it has not been widely adopted and relied on yet. Among machine learning methods, deep learning stands out in that it pertains to learning data representations automatically, as opposed to algorithms that require manual task-specific modifications. A convolutional neural network is a class of deep neural networks commonly applied to analyze 2-dimensional and 3-dimensional images, waveforms, and occasionally other sequential data such as text.

Convolutional Neural Networks (CNNs) were inspired by the structure and organization of the animal visual cortex, and neural *units* were modeled after neurons. A typical CNN consists of units organized in layers, with connections to units in adjacent layers. The network learns data representations from a *training* dataset, typically consisting of labeled data. A *test* dataset is used to evaluate performance, in which the trained network tries to predict the labels for the test data. When an input image is fed into the network during the training phase, each unit, associated with a weight, scans through parts of the image along with its adjacent units and generates corresponding output elements based on their weights. The aggregate of all outputs generated by scanning the complete input serves as the input for the next layer and the process continues. The weight associated with each unit is adjusted based on the difference between the generated output and the expected output at the final layer using a method known as backpropagation to better predict the expected output. This process is repeated for all images in the training dataset, with each unit storing a compounded weight over all the training images. However, when the network very accurately *memorizes* the training dataset, rather than *learns* to generalize from trends in it, a phenomenon known as overfitting can occur, resulting in poor performance on test data with high performance on training data. There are several ways to tackle this problem, but arguably the best is to use a larger well-represented training dataset. A well-represented dataset tries to mitigate sample biases from influencing the outcomes of the model. Once the model's performance is satisfactory, it is used to infer the output on new, unseen images. The authors previously worked with a CNN called Inception v3 to analyze renal biopsy slides and predict several outcomes, including the stage of chronic kidney disease, with success, in some metrics outperforming pathologists (based on their estimated fibrosis scores).

Inception v3 by Google is a widely used deep CNN in the computer vision community that

Correspondence: Adler J. Perotte, Department of Biomedical Informatics, Columbia University, New York, New York 10032, USA. E-mail: ajp2120@cumc.columbia.edu

has been shown to perform well on the ImageNet dataset, which contains millions of labeled images made available to the public. The units of Inception v3 are trained to recognize particular fundamental features, such as simple shapes and patterns that are subsequently used by other units within the model to recognize increasingly complex features ultimately accurately recognizing thousands of objects such as specific types of fruits, vehicles, animals, and musical instruments, for example. This hierarchical structure also allows the model to recognize patterns in other, completely different images, and is used to enable something called transfer learning: applying a model trained in one domain to another. The authors exploited this ability and further trained the model using biopsy images.

In addition, to generate a varied and rich training dataset, the authors performed data augmentation on existing biopsy images. Both this and transfer learning are strategies to mitigate the challenge of having few examples in a training dataset. Data augmentation is a process in which the training dataset is enriched, by performing classical image transformations like rotating, cropping laterally shifting and zooming existing images in the training dataset and adding them to it. The biopsy images used by the authors were quite large and computationally challenging to work with, so they cropped the image into smaller-sized images that could be processed with more ease. However, cropping also resulted in some glomeruli being chopped, and potentially missed by the CNN. The authors mitigated this by cropping with different offsets and integrating over the results. They also created new images using existing images by whitening a small fraction of the

pixels in the image. This enabled them to generate a more resilient model. The authors then divided this corpus into training and testing datasets. They repeated this process 4 times such that different images were used to train and test in each run. The objective here is to generate a well-represented training dataset while minimizing underlying biases that are not easily evident to humans, but could be picked up by the machine learning model to influence its decisions in an undesirable/unexpected way. The best performing model usually has minimal influence of these underlying biases, also known as a *well-generalized* model.

The model in the glomerular classification stage was used to predict 1 of 3 outcomes for each image: 0, no glomerulus detected; 1, normal or partially sclerosed glomerulus detected; and 2, globally sclerosed glomerulus detected. The coordinates of the cropped image in the original image was noted if the result was *2*. They then generated a heatmap of all the cropped images in which glomeruli were detected, across all different cropping methods, in the original image. This heatmap increased in brightness with the certainty of a glomerulus being present. This heatmap was further processed to mark the glomeruli with distinct boxes to highlight them. The trained model achieved an accuracy ranging from 89.66% to 95.06% over 4 independent training/testing runs, with kappas ranging from 0.8079 to 0.9111. The model that achieved the highest accuracy (most well-generalized model) was used for glomerular segmentation. The model identified nonglomerular regions with a specificity of 0.999 in the glomerular segmentation phase, but it is important to note that because most regions do not contain a glomerulus, this is a problem with

few positive regions and sensitivity/precision would better characterize performance.

In summary, the authors have used popular techniques from the image processing field and leveraged them to identify glomeruli in kidney biopsies from a racially and ethnically diverse cohort. They have leveraged transfer learning and popular regularization techniques to better generalize the model, like pixel whitening. They have also used other techniques, like cropping the image several times and aggregating over them. The glomerular classification models achieved an accuracy of 89.66% to 95.06%. The best performing model was selected for the glomerular segmentation phase, where it detected nonglomerular regions with a specificity of 0.999 and also marked and classified globally sclerosed glomeruli with a sensitivity of 0.558, F1-score of 0.623, and Matthew correlation coefficient of 0.628. This work demonstrates that deep learning models can assess complex histologic structures with high accuracy from digitized kidney biopsies.

DISCLOSURE

All the authors declared no competing interests.

REFERENCES

1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.

2. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J*. 2018;16:34–42.

3. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging*. 2016;35:1240–1251.

4. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network

surveillance of cranial images for acute neurologic events. *Nat Med.* 2018;24:1337–1341.

5. Liu V, Clark M, Mendoza M, et al. Automated identification of pneumonia in chest radiograph reports in critically ill patients. *BMC Med Inform Decis Mak.* 2013;13:90.

6. Kolachalama VB, Singh P, Lin CQ, et al. Association of pathological fibrosis with renal survival using deep neural networks. *Kidney Int Rep.* 2018;3:464–475.

7. Kannan S, Morgan LA, Liang B, et al. Segmentation of glomeruli within trichrome images using deep learning. *Kidney Int Rep.* 2019;4:955–962.