KI REPORTS
KIReports.org
TRANSLATIONAL RESEARCH

Association of Pathological Fibrosis With Renal Survival Using Deep Neural Networks

Vijaya B. Kolachalama[1,2,3,8], Priyamvada Singh[4,8], Christopher Q. Lin[5], Dan Mun[6], Mostafa E. Belghasem[7], Joel M. Henderson[7], Jean M. Francis[4], David J. Salant[4] and Vipul C. Chitalia[2,4]

[1]Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA; [2]Whitaker Cardiovascular Institute, Boston University School of Medicine, Boston, Massachusetts, USA; [3]Hariri Institute for Computing and Computational Science & Engineering, Boston University, Boston, MA, USA; [4]Renal Section, Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA; [5]College of Engineering, Boston University, Boston, Massachusetts, USA; [6]College of Health & Rehabilitation Sciences: Sargent College, Boston University, Boston, Massachusetts, USA; and [7]Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, Massachusetts, USA

Introduction: Chronic kidney damage is routinely assessed semiquantitatively by scoring the amount of fibrosis and tubular atrophy in a renal biopsy sample. Although image digitization and morphometric techniques can better quantify the extent of histologic damage, we need more widely applicable ways to stratify kidney disease severity.

Methods: We leveraged a deep learning architecture to better associate patient-specific histologic images with clinical phenotypes (training classes) including chronic kidney disease (CKD) stage, serum creatinine, and nephrotic-range proteinuria at the time of biopsy, and 1-, 3-, and 5-year renal survival. Trichrome-stained images processed from renal biopsy samples were collected on 171 patients treated at the Boston Medical Center from 2009 to 2012. Six convolutional neural network (CNN) models were trained using these images as inputs and the training classes as outputs, respectively. For comparison, we also trained separate classifiers using the pathologist-estimated fibrosis score (PEFS) as input and the training classes as outputs, respectively.

Results: CNN models outperformed PEFS across the classification tasks. Specifically, the CNN model predicted the CKD stage more accurately than the PEFS model ($\kappa = 0.519$ vs. 0.051). For creatinine models, the area under curve (AUC) was 0.912 (CNN) versus 0.840 (PEFS). For proteinuria models, AUC was 0.867 (CNN) versus 0.702 (PEFS). AUC values for the CNN models for 1-, 3-, and 5-year renal survival were 0.878, 0.875, and 0.904, respectively, whereas the AUC values for PEFS model were 0.811, 0.800, and 0.786, respectively.

Conclusion: The study demonstrates a proof of principle that deep learning can be applied to routine renal biopsy images.

Kidney Int Rep (2018) **3,** 464–475; https://doi.org/10.1016/j.ekir.2017.11.002
KEYWORDS: histology; machine learning; renal fibrosis; renal survival

From self-driving cars to face recognition, artificial intelligence and machine learning (ML) algorithms are being widely applied to enhance human endeavors. Over the last few years, the scientific community has witnessed a rapid increase in the adoption of cutting-edge data analytic tools such as ML to address several questions in clinical medicine.[1–6] ML techniques give computers the ability to integrate discrete sets of data in an agnostic manner to find hidden insights and to generate a disease-specific fingerprint. These tools are now being rapidly adopted in several specialties as unbiased, self-learning approaches for pathologic assessment. Such a framework can leverage hundreds to thousands of images as inputs and allow for objective quantification, followed by their association with several clinical outcomes of interest. ML techniques also have the potential to uncover several nonintuitive features that may be clinically relevant and hypothesis generating, as demonstrated in other disease scenarios.[7]

Although the trained eyes of expert pathologists are able to gauge the severity of disease and to detect

Correspondence: Vijaya B. Kolachalama, Boston University School of Medicine, 72 East Concord Street, Evans 636, Boston, Massachusetts 02118, USA. E-mail: vkola@bu.edu

[8]The first 2 authors contributed equally to this work.

Received 7 August 2017; revised 4 October 2017; accepted 6 November 2017

nuances of histopathology with remarkable accuracy, such expertise is not available in all locations, especially at a global level. Moreover, there is an urgent need to standardize the quantification of pathological disease severity, such that the efficacy of therapies established in clinical trials can be applied to treat patients with equally severe disease in routine practice. The tools to do this are at hand in the form of digitized images of pathology sections prepared with routine stains and ML algorithms. Such methods have already been tested and shown to be reliable for the analysis of malignancies such as cancer.[7–17] The application of deep learning frameworks, such as convolutional neural networks (CNN) for object recognition tasks, is proving to be especially valuable for classification of several diseases.[8–29]

To test the feasibility of applying ML technology to the analysis of routinely obtained kidney biopsy samples, we performed a proof-of-principle study on kidney biopsy sample sections with various amounts of interstitial fibrosis as revealed with Masson trichrome stain (Figure 1). Using an established CNN that relies on pixel density of digitized images,[30] we analyzed the ability of the ML technique to quantify the severity of disease as determined by several clinical laboratory measures and renal survival (Supplementary Figure S1). CNN model performance was then compared with that of the models generated, using the amount of fibrosis reported by an expert nephropathologist as the sole input and corresponding laboratory measures and renal survival as the outputs.

METHODS

Data Collection

A retrospective analysis of renal biopsy findings was performed on patients treated at the Boston Medical Center (BMC) between January 2009 and December 2012. Reports from all follow-up visits between 2009 and 2016 for these patients were also reviewed. All patient data were collected under protocol H-32289, which was reviewed and approved by the Institutional Review Board at Boston University Medical Campus. More than 300 biopsy samples were processed at BMC, of which 171 biopsy slides were available for subsequent imaging. These biopsy samples were obtained from adult patients who had 1 or more native or renal allograft biopsies, independent of the indication for the biopsy procedure. The only criterion for inclusion was the availability of pathological slides and accompanying clinical data. Several demographic and clinical features (including estimated glomerular filtration rate [eGFR], baseline creatinine, nephrotic-range proteinuria, etc.) were collected on these patients at the time of biopsy (Table 1). The 4-parameter Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula was used to calculate eGFR. A detailed chart review of the patients' electronic medical record was used to estimate 1-, 3-, and 5-year renal survival (see Supplementary Material for details of clinical data collection). Renal survival was measured as the time from the day of biopsy until the patient had 1 of the following events: initiation of dialysis, renal transplant, or all-cause mortality.

Imaging

Kidney biopsy samples were obtained in the form of individual trichrome-stained slides. Each selected core was imaged at ×40, ×100, and ×200 magnifications using a Nikon Eclipse TE2000 microscope (Melville, NY; http://www.bumc.bu.edu/busm/research/cores/). For ×40

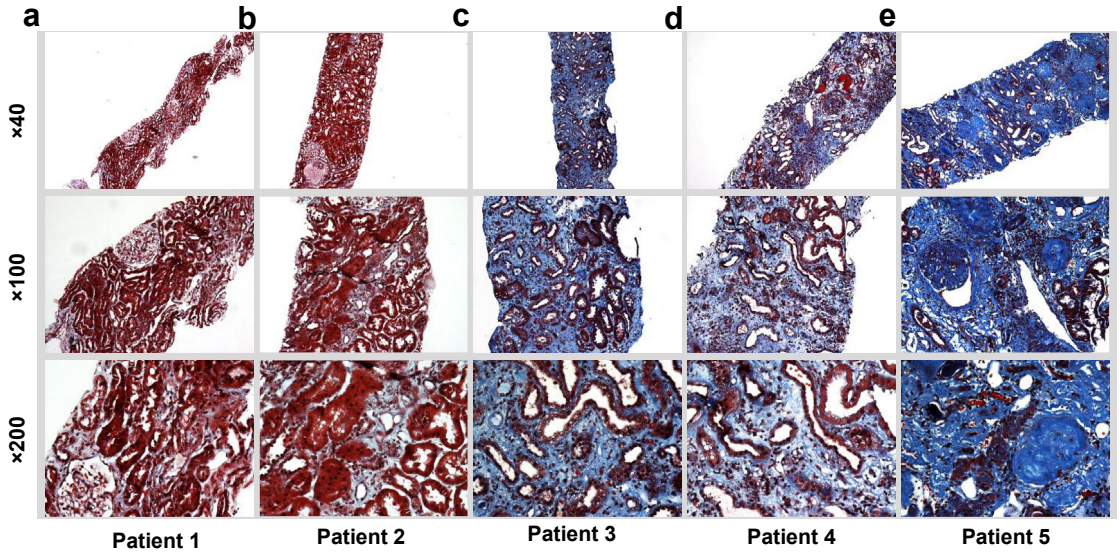


Figure 1. Sample interstitial fibrosis cases from the patient cohort. The trichrome-stained images demonstrate the variability and extent of interstitial fibrosis observed within renal biopsy samples at different magnifications. The in-house nephropathologist–derived fibrosis score was 5% to 10% for (a), 20% for (b), 30% for (c), 50% for (d), and 85% for (e).

Table 1. Characteristics of study population[a]

Characteristic	Value	Units
Number of patients	171	—
Age, median (range)	52 (19–86)	yr
% Male	59.6	%
Number of patients per race/ethnicity (white, black, Hispanic, other)	46, 79, 24, 22	—
BMI, median (range)	28.94 (15 –56.2)	kg/m^2
Native kidney biopsy samples	110	
Creatinine, median (range)	2.31 (0.54–13.29)	mg/dl
eGFR, median (range)	30 (5–163)	ml/min per 1.73 m^2
Proteinuria, median (range)	1.79 (0.03–20.5)	g/g
Interstitial fibrosis, median (range)	30 (0–90)	%
% Patients with 1-yr renal survival	79.1	%
% Patients with 3-yr renal survival	58.9	%
% Patients with 5-yr renal survival	39.3	%

BMI, body mass index; eGFR, estimated glomerular filtration rate.
[a]A team of nephrologists performed a detailed chart review to extract demographic, biopsy, and other clinical data from patients who underwent treatment for chronic kidney disease at the Boston Medical Center between 2009 and 2012. The task also included detailed chart review of follow-up reports between 2009 and 2016 for these patients.

magnifications, images were generated with a special consideration to cover the entirety of the sample. This usually resulted in a minimum of 3 images and a maximum of 14 images, with the majority of the samples requiring 6 images to fully capture the full length of the core. For ×100 and ×200 magnifications, about 5 images per magnification were taken sequentially from 1 end of the sample to the other, with almost no overlapping regions between images. All of the images were manually focused using the NIS-Elements AR software (Nikon, Tokyo, Japan) that was installed on the computer connected to the microscope.

Model Training

We used Google's Inception v3 architecture pretrained on millions of images with 1000 object classes by incorporating minor changes to fine-tune the framework and to associate trichrome image features with the clinical phenotypes[29,30] (Figure 2a). Specifically, we removed the final classification layer from the network and retrained it with our dataset using the training classes defined based on the problem of interest (i.e., CKD stage, serum creatinine, and nephrotic-range proteinuria at the time of biopsy, and 1-, 3-, and 5-year renal survival). We then performed fine-tuning of the parameters at all layers. During training, we resized each image to 299 × 299 pixels to make it compatible with the original dimensions of the Inception v3 network architecture and leveraged the image features learned by the ImageNet pretrained network.[30] This procedure, known as transfer learning, is optimal, given the amount of data available.

Our deep neural network was trained using back-propagation.[31] Using the framework of transfer learning, we first trained only the top layers that were randomly initialized by freezing all the convolutional layers. We then trained the model using our data for several epochs with the early stopping criteria that monitored the validation loss. We used "rmsprop" with a decay of 0.9, momentum of 0.9, and epsilon of 0.1, along with the use of L1, L2 hybrid regularizers (0.01, 0.01). After the top layers were trained, we performed fine-tuning of the convolutional layers by freezing the bottom (N) layers and training the remaining top layers. Sensitivity analysis was performed to identify the optimal layer (N) for each problem under consideration (Supplementary Figure S2a and b). We then recompiled the model for the above modifications to take into effect. For this task, we used stochastic gradient descent with a learning rate of 0.001 with momentum of 0.9, and L1 and L2 regularizers (0.01, 0.01). We finally trained the model again by fine-tuning the top blocks alongside the top dense layers. We used Google's TensorFlow (https://www.tensorflow.org) back-end to train, validate, and test our network.

During training, images were augmented by factors of 10 or 100, depending on the problem of interest. Each image was rotated randomly between 0° and 180°, randomly shifted in the horizontal and vertical directions, shifted the channels, sheared, and zoomed, all by a scale of 0.1. Images were also flipped vertically and horizontally, with a probability of 0.5. All these aspects are part of a well-known strategy called data augmentation.[32–37]

The first modeling task was a multilabel classification problem in which the resized trichrome images were used as the input and the eGFR-based CKD stage (stages 1–5) defined using the Kidney Disease Outcomes Quality Initiative (KDOQI) guidelines at the time of biopsy was used as the output (Figure 2b). The rest of the tasks were binary classification problems in which the resized trichrome images remained as the input, and binarized values of baseline creatinine >1.3 mg/dl for men and >1.1 mg/dl for women) and nephrotic-range proteinuria (>3.5 g/d), as well as 1-, 3-, and 5-year renal survival values served as outputs, respectively (Figure 2b).

Training and testing of all the models were performed on a 14-core 2.4 GHz Intel Xeon E5-2680v4 processor with Broadwell CPU architecture and an NVIDIA Tesla P100 GPU card with 12 GB of memory that is located on the Boston University's Shared Computing Cluster (https://www.bu.edu/tech/support/research/computing-resources/scc/).

Performance Metrics

For the binary classification problems (gender-specific high/low creatinine, presence or absence of

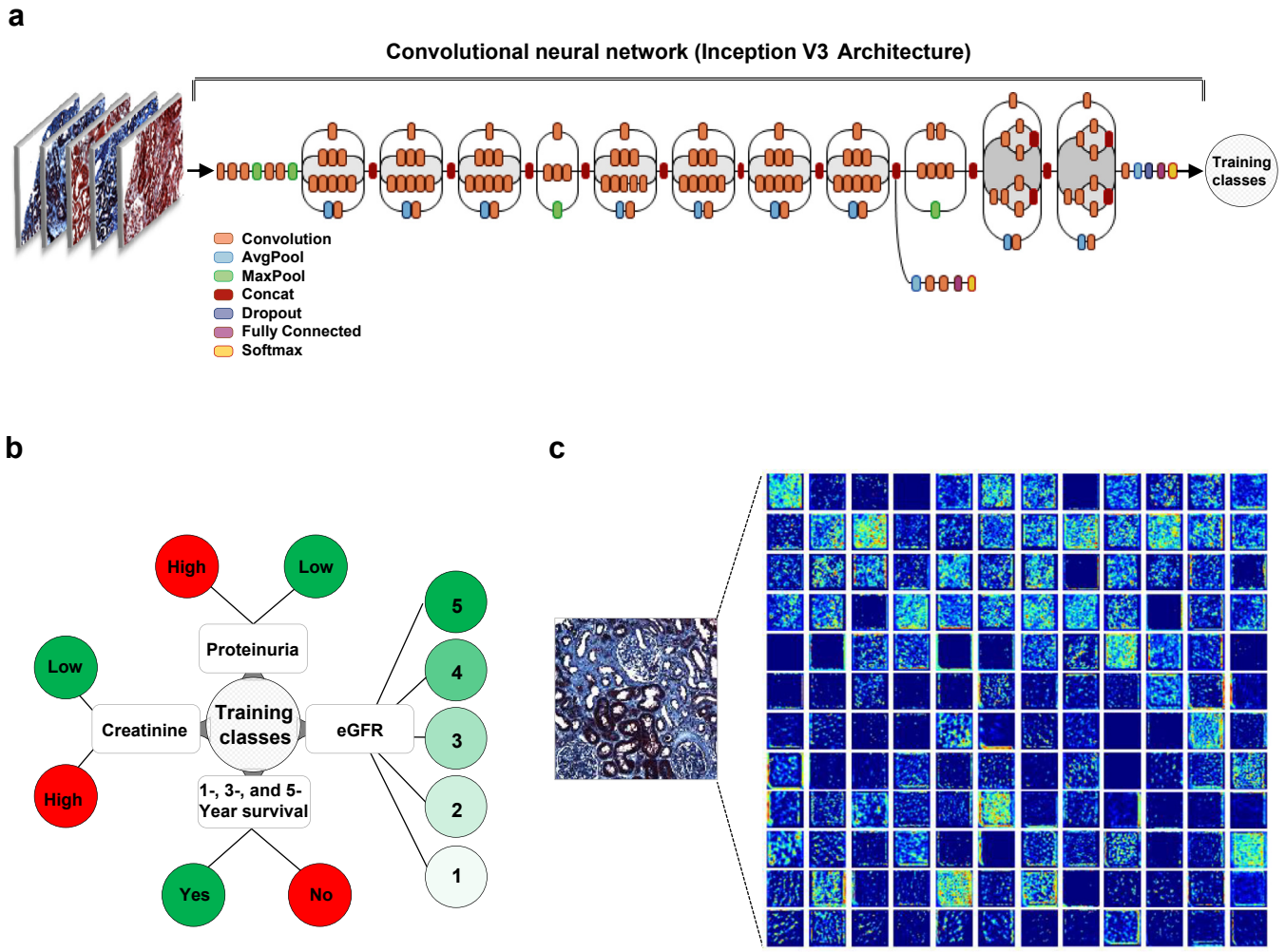


Figure 2. Deep neural network model. (a) Our classification technique is based on using a transfer learning approach on Google Inception V3 convolutional neural network (CNN) architecture pretrained on the ImageNet dataset (1.28 million images over 1000 generic object classes) and fine-tuned on our dataset (see Methods). Inception v3 CNN architecture reprinted with permission from the Google blog "Train Your Own Image Classifier With Inception in TensorFlow" (https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html). (b) Using the dataset containing trichrome-stained images from the patients as inputs, several models were constructed with different output classes (chronic kidney disease stage based on estimated glomerular filtration rate [eGFR], binarized serum creatinine and proteinuria, as well as 1-, 3-, and 5-year renal survival). (c) Visualization of filters generated during training. Only 144 of the 256 filters used at the first pooling layer are shown.

nephrotic-range proteinuria, and 1-, 3- and 5-year renal survival), the performance of the deep neural network was computed using the c-statistic or area under curve (AUC) computed on the receiver operating characteristic (ROC) curve. We also computed F1 score as a measure of model accuracy that considers both the precision and recall of a test. It is defined as follows:

$$F1 = 2 * \mathrm{TP}/(2 * \mathrm{TP} + \mathrm{FP} + \mathrm{FN}). \quad \text{(Equation 1)}$$

Here, TP denotes true-positive values, and FP and FN denote false-positive and false-negative cases, respectively (see Supplementary data for more details).

We also computed Matthews correlation coefficient (MCC),[38] which is a balanced measure of quality for dataset classes of different sizes of a binary classifier and defined as follows:

$$MCC = [(\mathrm{TP} * \mathrm{TN}) - (\mathrm{FP} * \mathrm{FN})]/[(\mathrm{TP} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})]^{0.5}. \quad \text{(Equation 2)}$$

Here, TN denotes true negative values. For the multilabel classification problem (i.e., prediction of eGFR-based CKD stage), model accuracy and Cohen's κ score were computed.[39–41] The κ statistic measures interrater agreement for categorical items. Note that accuracy of 1 is a perfect score, and that a κ score of 1 indicates perfect agreement between the true and predicted labels.

RESULTS

Patient Population

Since this was a retrospective study, we had the advantage of carefully curating all the digital images as well as extracting corresponding baseline characteristics and clinical phenotypes for each patient. The baseline characteristics are representative of a population of patients attended to at a nephrology service at a teaching hospital (Table 1). About 60% patients were male, and about 46% of the overall population was African American. The majority of patients had hypertension (~84%) and cardiovascular disease (~75%), and about 43% had diabetes. About 64% of the biopsies were from native kidneys, and the remainder were from transplanted kidneys. About 82% of patients had CKD stage 3 to 5 based on the eGFR using the CKD-EPI formula; 6% had stage 2 CKD, and the rest had stage 1 CKD. About 35% of patients had nephrotic-range proteinuria (>3.5 g/d).

Imaging

Light microscopy images of Masson trichrome–stained sections were captured at 3 different magnifications (×40, ×100, and ×200) and processed in NIS-Elements AR software (Nikon) (Figure 1). In total, we processed kidney biopsy from 171 patients, which resulted in 2255 unique images that were saved in 16-bit TIFF format. These images were then converted to 8-bit red–green–blue (RGB) color images in TIFF format, resized, and subsequently used for training the deep neural network (Figure 2).

Performance Metrics

We developed 6 independent CNN models to better associate information derived from the trichrome images with the clinical phenotypes (CKD stage based on eGFR, creatinine, and nephrotic-range proteinuria, all at the time of biopsy) and renal survival (Figure 2b). For each classification task, we also developed a model to associate pathologist-derived fibrosis score (PEFS) with the output class of interest that served as the baseline for comparison.

For each task, a detailed set of sensitivity analysis was performed, including the layer selection for fine-tuning as well as the impact of batch size on model performance. The CNN model with the best cross-validated performance was selected. For the eGFR model, model accuracy and κ values were computed.[42] For the binary classification models, area under the curve (AUC or c-statistic), F1 score (Equation 1 in the Methods), and Matthews correlation coefficient (MCC) (Equation 2 in the Methods) were computed (Supplementary Figure S3). Note that the F1 score is a widely used metric in the ML community, as it considers both the precision and the recall of a test, and thus it can be viewed as a weighted average of precision and recall. However, the F1 score does not take the true negatives into account, and therefore composite metrics such as MCC can be considered as more robust measures of performance. In essence, MCC is a correlation coefficient between the observed and predicted binary classifications and a balanced measure that can be used even if the classes are of different sizes.[38]

Nephropathologist Model Based on Fibrosis Score

Our first goal was to develop baseline models by which to compare the performance of the CNN models. Because PEFS was the input, a linear discriminant classifier was used to associate this score with the respective output class labels (CKD stage, serum creatinine, etc.). Model results indicated that the performance of the pathologist model was remarkably good, considering that only a single scalar value derived from detailed visual assessment by the nephropathologist was used as the input feature (Figures 3–5). This result underscores the notion that fibrosis score as such is an important predictor of clinical phenotypes including renal survival.

Also, choosing a different learning algorithm for the classification task of associating PEFS with a clinical phenotype did not alter our overall inferences related to the comparison with the CNN model performance. We used 3 different algorithms—linear discriminant analysis, Naïve Bayes, and support vector machine classifiers—to build several classifiers with PEFS as the input and different clinical phenotypes (serum creatinine and proteinuria at the time of biopsy as well as 1-, 3-, and 5-year renal survival) as the outputs, respectively. Model training, followed by validation on test data that had not been used for training, resulted in similar performances of the binary classifier (Supplementary Figure S4).

CNN Model for Prediction of Clinical Phenotypes at the Time of Biopsy

Our next task was to evaluate the predictive capability of the CNN model. The CNN models developed using the transfer learning approach outperformed the models derived using PEFS on all the classification tasks. Three separate CNN models were trained to associate the images with the stage of kidney disease at the time that the biopsy procedure was performed. The first model was related to predicting CKD stage based on eGFR at the time of biopsy, and our deep learning model resulted in a superior performance ($\kappa = 0.519$ [CNN model] vs. $\kappa = 0.051$ [pathologist])

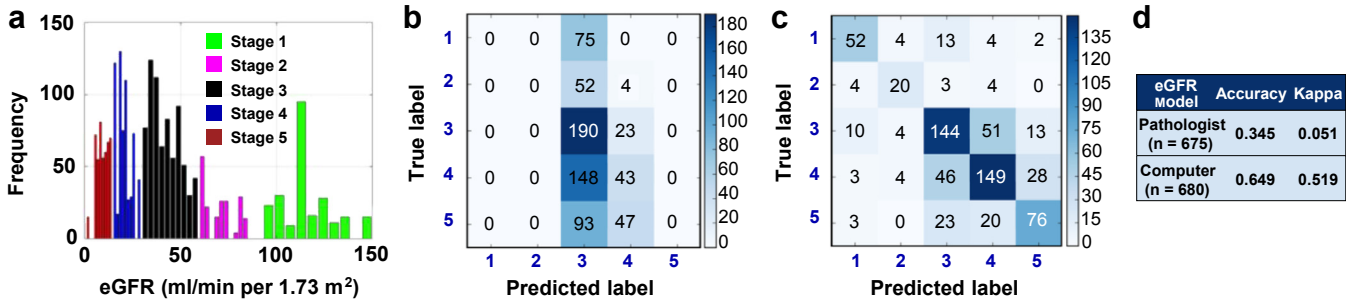


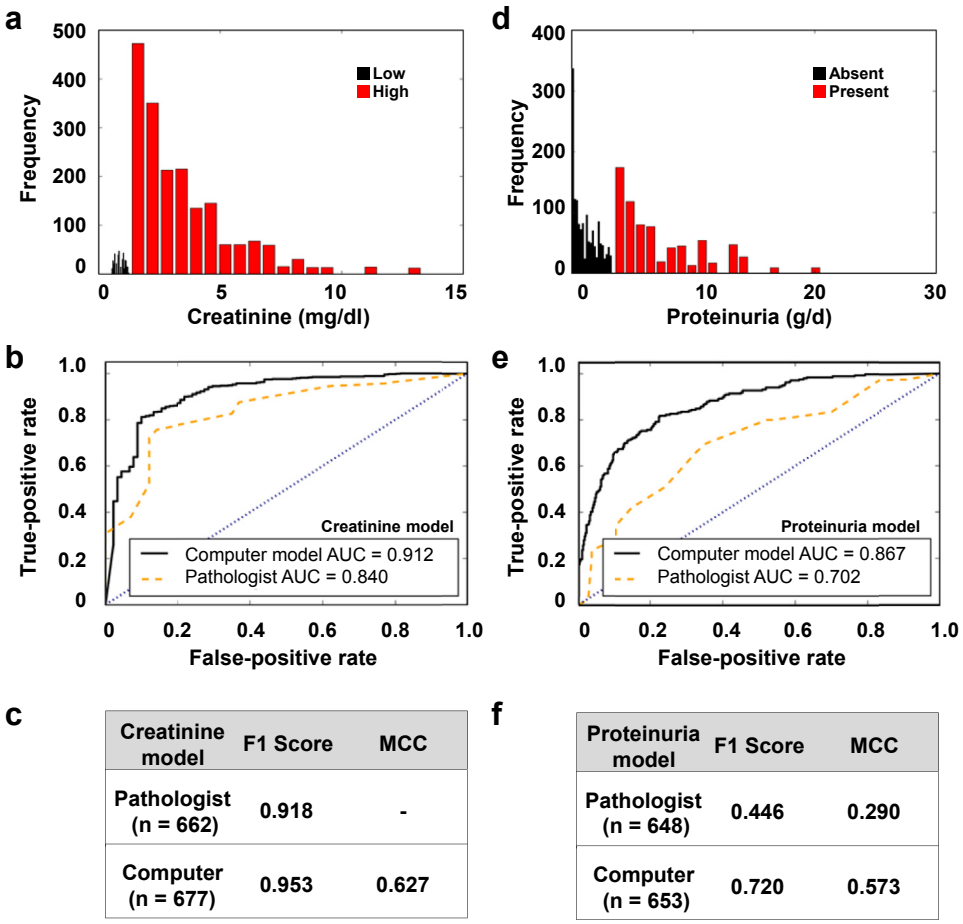
Figure 3. Predictive model of estimated glomerular filtration rate (eGFR) at the time of biopsy. (a) Distribution of eGFR values across the patient cohort. The histogram frequency corresponds to the number of images. (b) A multilabel linear discriminant classifier was trained on the data with pathologist-derived fibrosis as the input and eGFR-based chronic kidney disease (CKD) stage (stages 1–5) at the time of biopsy as the output. Image data were randomly split such that 70% of the data (n = 1512) were reserved for model training and the remaining for testing (n = 648). "True label" denotes the CKD stage derived from calculated eGFR values at the time of biopsy, whereas "Predicted label" indicates the model assessment of the CKD stage. (c) Fine-tuned convolutional neural network (CNN) model was used to predict on test image data (n = 677) not used for training. Performances of the pathologist (b) and the CNN (c) models are shown in the form of confusion matrices. (d) A κ score was computed by comparing model-derived output values with the clinically reported values of eGFR. The CNN model accuracy and κ score indicate the superior performance of the CNN model in comparison to the pathologist model.

(Figure 3b–d). The CNN architecture resulted in superior performance even for models constructed using images processed at only a single magnification. For example, we trained a separate CNN model using images taken at ×100 magnification and predicted on test data of ×100 images not used for training. For comparison purposes, we extracted PEFS from the clinical biopsy reports for these images, and constructed a classifier using linear discriminant analysis. Even for this case, the CNN model outperformed the model derived using PEFS (Supplementary Figure S5).

CNN models used for the binary classification tasks of predicting clinical phenotypes such as gender-specific high/low value of creatinine as well as nephrotic-range proteinuria also outperformed the performance of the models constructed using PEFS. Specifically, for the creatinine CNN model, the c-statistic (or AUC) was 0.912 versus an AUC of 0.840 for the model derived using PEFS (Figure 4b). The superiority of the CNN model is exemplified with the help of other metrics including F1 score (0.918 [pathologist)] and 0.953 [model]), and MCC value of 0.627 for the CNN model (Figure 4c). Note that MCC computation for the case of the pathologist model did not result in a numeric value because both the false-positive and true-negative values were 0 (Equation 2). Similarly, the CNN model predicted the presence of nephrotic-range proteinuria (AUC 0.867) more accurately than the model derived using PEFS (AUC 0.702) (Figure 4e). Other metrics that were computed for this case also showed the superiority of the CNN model (F1 score: 0.446 [pathologist] vs. 0.720 [CNN model]; MCC: 0.290 [pathologist] vs. 0.573 [CNN model]) (Figure 4f). Although we recognize that one does not need a kidney biopsy to determine the stage of CKD, serum creatinine level, or presence of nephrotic-range proteinuria, this exercise allowed us to test the CNN model's predictive ability against hard numeric (or categorical) values and to proceed with the task of testing its prognostic value.

Prognostic Value of the CNN Model

One of the most interesting aspects of leveraging the CNN model architecture was to evaluate its ability to train the biopsy images for prognostic purposes. Specifically, we trained independent CNN models for 3 binary classification tasks focused on predicting 1-, 3-, and 5-year renal survival, respectively. All patients who underwent a kidney biopsy and were followed up respectively for 1, 3, and 5 years after the biopsy were included in the study. Specifically, kidney biopsy data as well as 1-year renal survival information were available on 158 patients (2050 images), followed by biopsy data and 3-year survival available on 146 patients (1900 images), and finally biopsy data and 5-year survival information available on 117 patients (1517 images). All 3 models resulted in better performance than the respective models developed using the PEFS alone. Specifically, the AUC values for the CNN model for 1-, 3-, and 5-year renal survival were 0.878, 0.875, and 0.904, respectively (Figure 5a–c), whereas the AUC values of the model developed using PEFS for 1-, 3-, and 5-year renal survival were 0.811, 0.800, and 0.786, respectively (Figure 5a–c). Similarly, other metrics such as F1 score and MCC computed for these models also confirmed the superiority of the CNN model over the classification model generated using PEFS (Figure 5d–f). Thus, the CNN architecture accurately reflects the stage of kidney disease as well as renal survival.



c

Creatinine model	F1 Score	MCC
Pathologist (n = 662)	0.918	-
Computer (n = 677)	0.953	0.627

f

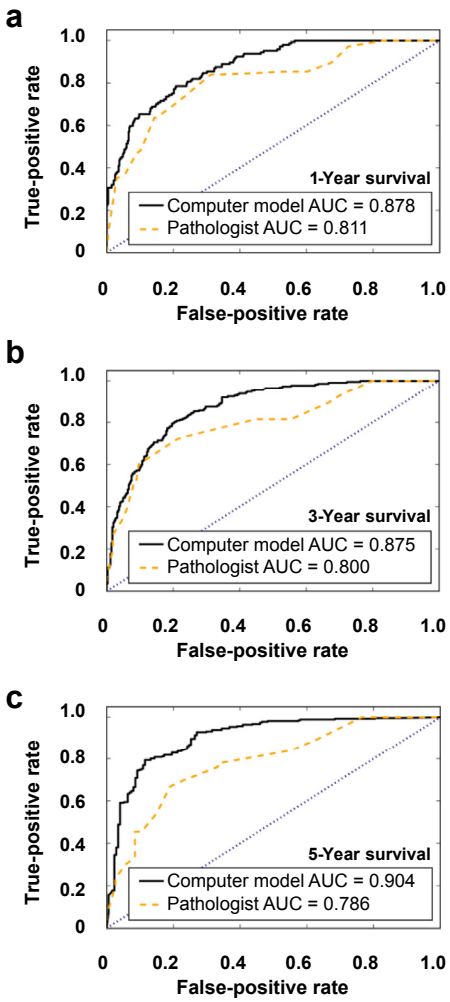
Proteinuria model	F1 Score	MCC
Pathologist (n = 648)	0.446	0.290
Computer (n = 653)	0.720	0.573

Figure 4. Predictive models of creatinine and nephrotic-range proteinuria at the time of biopsy. (a) Distribution of creatinine values across the patient cohort. The histogram frequency corresponds to the number of images. Color to a set of bars within the histogram was assigned based on the Kidney Disease Outcomes Quality Initiative (KDOQI) guideline–driven cutoff values for high and low creatinine. (b) A binary linear discriminant (BLD) classifier was trained using 70% of the image data (n = 1545), with pathologist-derived fibrosis value as the input and baseline creatinine value at the time of biopsy as the output. Model predictions were performed on the remaining 30% of the data (n = 662), and a receiver operating characteristic (ROC) curve was generated. (c) Both F1 score and Matthews correlation coefficient (MCC) computed using models' performances on test data indicate superior performance of the convolutional neural network (CNN) model. (d) Distribution of proteinuria values across the patient cohort. Color to a set of bars within the histogram was assigned based on the KDOQI guideline–driven cutoff value for nephrotic-range proteinuria (g/d). (e) A similar BLD classifier was trained using 70% of the image data (n = 1512), with the pathologist-derived fibrosis value as the input and the clinical indication of proteinuria as the output. Model predictions were performed on the remaining 30% of the data (n = 648), and an ROC curve was generated. (f) Both the F1 score and the MCC computed using models' performances on test data indicate superior performance of the CNN model in comparison to the pathologist model.

DISCUSSION

Renal biopsy is an invasive and yet an invaluable procedure for the diagnosis, prognosis, and treatment of patients with kidney disease.[43] Detailed histologic processing of biopsy samples yields important pathologic information, and digitization of these specimens lends itself to the possibility of using computer-based technologies for better quantification. In this study, we chose to apply ML technology to the analysis of interstitial fibrosis, a common manifestation of a wide variety of kidney diseases and a prognostic indicator of chronic kidney disease progression in humans.[44–47] An important strength of the study is that the ML technology was applied to trichrome-stained histologic images of routine kidney biopsy samples without any special processing or manipulation other than digital scanning, which allowed us to directly compare the results of the ML analysis with those derived from the clinical pathological report on the same specimens. Although our results demonstrating the utility of ML are limited to the analysis of interstitial fibrosis, the developed ML framework can be extended to associate other complex renal pathologies with several clinical phenotypes.

Deep learning algorithms such as CNN, powered by advances in hardware as well as software and well-curated datasets, have recently been shown to exceed human performance in visual object recognition tasks.[31] Although there is an increasing focus on



d

1-Year survival	F1 score	MCC
Pathologist (n = 610)	0.477	0.397
Computer (n = 615)	0.641	0.552

e

3-Year survival	F1 score	MCC
Pathologist (n = 564)	0.695	0.538
Computer (n = 570)	0.748	0.580

f

5-Year survival	F1 score	MCC
Pathologist (n = 450)	0.798	0.435
Computer (n = 455)	0.848	0.605

Figure 5. Predictive models of 1-, 3-, and 5-year renal survival. Three separate binary linear discriminant classifiers were trained using 70% of the image data, with pathologist-derived fibrosis value as the input and 1-, 3-, and 5-year renal survival values computed from the clinical reports as the outputs, respectively. The convolutional neural network (CNN) framework was also used to train separate models with the 3 outputs. Predictions of the models were performed on the remaining 30% of the data, denoted as "n" for each case. Receiver operating characteristic curves comparing the pathologist model with the CNN model were generated for each case ([a] 1-year, [b] 3-year, and [c] 5-year renal survival) respectively. F1 score and Matthews correlation coefficient (MCC) values computed for each case ([d] 1-year, [e] 3-year, and [f] 5-year renal survival) on the test data also indicate superior performance of the CNN model.

developing these approaches using several modalities of data, such as audio, text, and video, our work focused specifically on image-based classification. The CNN architecture has been shown to be optimal for these tasks, whereas other deep learning methods can be used for applications above and beyond visual object recognition.[31] Moreover, we are witnessing an increasing interest in image-based classification of a clinical phenotype using deep learning, and it has been shown to be extremely effective in several clinical disciplines such as dermatology, radiology, and oncology.[7–9,13,15–18,20,21,23,25,28,29,48–51] To the best of our knowledge, no body of work related to applications of deep learning in nephropathology has been published to date. In this article, we have outlined the development of a CNN that at least matches the performance of a skilled nephropathologist's ability to quantify the extent of kidney fibrosis as it relates to 6 key classification tasks: determination of CKD stage based on eGFR, baseline serum creatinine and nephrotic-range proteinuria at the time of biopsy, as well as 1-, 3-, and 5-year renal survival. Although the case with eGFR as the output was modeled as a multi-label classification task (CKD stages; 5 labels), the rest were modeled as binary classification tasks.

As part of routine clinical workflow, the expert nephropathologist carefully reviews the trichrome stain–based histological images and assigns a grade of renal fibrosis. Typically, the grading is discrete and ranges from 0 to 100, with 0 assigned to no detectable fibrosis and 100 assigned to the biopsy with complete fibrosis. Such a derived grading process resulting in a single scalar value has shown to have remarkable clinical utility.[43,46,47] This was confirmed in this study

by the models that we developed using the pathological grading as the input and the output classes derived from a detailed chart review of the clinical reports spanning more than several years (Figures 3–5).

To evaluate the performance of the CNN model on the same biopsy samples, we performed a computer-based analysis of the digitized biopsy images and related the findings to the outcomes of interest. This task resulted in a dataset comprising 171 patients with 2255 images that were subsequently used for model training. As shown in Figures 3 to 5, training and testing on these images, followed by predictions on test data, demonstrated superior CNN model performance in comparison to the linear discriminant classifier trained using the pathologist-derived fibrosis score (Supplementary Table S1).

Superiority of the CNN model can be attributed to several factors. First, the concept of transfer learning has boosted the use of CNN models for classification tasks with smaller datasets, as they have begun to single-handedly disconnect the presumed link between larger dataset size and high model accuracy. This is because we can use pretrained CNN models with features that were already learned, using millions of images with different object classes, and that produced outstanding results exceeding human performance.[29] Moreover, transfer learning usually results in faster training times than required for constructing a new CNN from scratch, as there is no need to estimate all the parameters in a pretrained CNN. Despite these advantages, we still used a rigorous cross-validation strategy of splitting the dataset into 2 portions and validating our CNN model using the test data that were not used for model training. The transfer learning approach turned out to be very effective, given the number of trichrome images that were generated from 171 patients. Also, to make the most of this dataset, we performed "data augmentation" whereby each image was subjected to random transformations and then fed into the CNN model as multiple images for training.[31] Moreover, the process of introducing additional variability/noise over that which could possibly be contained within the original images has been shown to limit model overfitting and to increase model generalizability.[31–37]

The second reason that explains the superior CNN performance is related to the deep neural network architecture itself. Using operators such as convolution, activation, and pooling (or subsampling), training a CNN model involves performing these operations multiple times in a systematic fashion to transform pixel-level information to high-level features of the input image. These features are then used to perform the classification task using fully connected layers along with training the model using another process called backpropagation, a method that is widely used in the ML community to train artificial neural networks.[31] Therefore, the CNN model training is in stark contrast to the pathologist model training that was done using a single value (fibrosis score) as an input feature along with corresponding output classes. This aspect underscores the value of leveraging a computer algorithm such as CNN to capture pixel-level information derived from the whole image and to associate it with an outcome of interest but not a fibrosis score *per se*.

Even though the CNN model outperformed on all the classification tasks, the model derived using PEFS was a close second. This was especially the case with the creatinine, proteinuria, and survival models, as the performances were almost comparable, with the CNN model slightly outperforming the PEFS model each time. However, for predicting the CKD stage, the CNN model resulted in a much better accuracy (Figure 3 and Supplementary Figure S5). One possible explanation is that predicting the CKD stage is a multilabel classification problem with 5 different outcomes, whereas all the other cases are binary classification problems with only 2 different outcomes. Because the PEFS model has only a single scalar value for training, one could argue that CNN model training that involves fine-tuning of several parameters is more suitable for tackling both binary and multilabel classification problems.

Accurate prediction of renal survival and other clinical phenotypes can inform therapeutic interventional strategies. CNN model architectures are beginning to address this aspect with a high level of success, and therefore one can envision the utility of these tools within the daily clinical workflow. A strength of this study is that it demonstrates the ability of an ML algorithm applied to digital images of routinely prepared kidney biopsy slides to assess the severity of pathology and its relationship with clinical measures with accuracy at least equivalent to that of an experienced nephropathologist without the need for labor-intensive morphometric analysis by a skilled observer. Although similar accuracy can be achieved, as recently documented, by meticulous scoring of pathological elements,[43] such methods are not applicable to routine anatomic pathology.

If the CNN architecture is embedded in the form of software application (or "app") within a mobile device or an easily accessible computer, then predictions based on the digitized histologic images can be derived. Using our data as an example, a potential clinical workflow could involve the pathology clinic receiving a kidney biopsy sample, and processing it using trichrome or other staining protocol, followed by sample digitization into a red–green–blue (RGB) color image.

If this color image were directly fed into the app containing the trained CNN model installed on the pathologist's computer, then predictions of 1-, 3-, and 5-year renal survival could be produced as outputs of the model instantaneously. These predictions would provide added value to the biopsy findings and, along with other clinical assessments, would enable a more precise care management and follow-up strategy on the biopsied patient. This technology could become all the more useful in areas where nephropathology expertise is not readily available, so a general pathologist could directly interpret these images using the app and could prepare the clinical reports accordingly to facilitate care management. For all these cases, prospective validation is needed in order for the app to be considered as a clinical tool. One could imagine designing a prospective study in which the app driven by the CNN model predicts the outcome (such as 5-year renal survival) immediately after kidney biopsy is performed on a patient. We can then track the patient over time and check whether the model prediction validates with the actual time of renal survival.

The ML algorithms clearly have limitations and provide incremental value rather than replacing the human factor. We acknowledge that a nephropathologist's clinical impression and diagnosis is based on contextual factors above and beyond visual and pathologic inspection of a lesion in isolation. Nevertheless, the ability to classify histologic images using a computer with the accuracy of an experienced nephropathologist has the potential to affect renal practice, especially in resource-limited settings.

Another limitation of our study is that we did not consider the possible impact of treatment on renal survival. For example, it is possible that treatment with immunosuppressive drugs might have altered the course and influenced renal survival, which of course, could not have been predicted by ML analysis at the time of the original biopsy. In other words, renal survival might have been better in such cases than our ML model would have predicted. Despite that, the model was remarkably good at predicting survival, and one can only assume that the extent of kidney fibrosis is a major determinant of outcome regardless of treatment. Thus the value of the ML analysis in such cases would be to exclude cases in which treatment would be futile and in which the risks would outweigh the benefits.

Our work relied only on images derived using the Masson trichrome stain, as it has long been used as a standard for quantifying fibrosis not only in renal but in other organ pathologies. Some pathologists, however, rely on other protocols such as Sirius Red staining, and recent work in this area has demonstrated staining protocol–specific accuracies in terms of quantifying fibrosis.[51] The CNN model described here could likely be applied equally well to images generated with Sirius Red or other staining protocols, including periodic acid–Schiff. One could also imagine extending the CNN architecture to model image data derived directly from whole slide imaging. The whole slide imaging–derived images are typically of higher-bit graphics (~16 bits), and thus require special handling and pre-processing even before they can be used for model construction.[43]

Although we performed extensive studies to identify the best batch size, learning rate, augmentation factor, and CNN layers for network fine-tuning of images stained to highlight fibrosis, this technology lends itself to the analysis of several other parameters such as glomerular sclerosis and immunohistochemistry of inflammatory cell infiltrates or specific cell types or components. Furthermore, the CNN framework allows for the development of glomerular disease–specific models to evaluate focal segmental glomerulosclerosis and chronic allograft nephropathy, and possibly to identify more accurately those cases in which immunosuppressive therapy is likely to be futile because of the extent of kidney damage.

In conclusion, in this study, we demonstrated the effectiveness of using deep learning architecture in nephropathology, a technique that enabled us to associate trichrome-stained histologic images obtained at the time of biopsy with several clinical phenotypes. Using a single CNN architecture trained on these images, we compared the CNN performance with that of an experienced nephropathologist by testing the model across 6 classification tasks: CKD stage based on eGFR, high/low creatinine and nephrotic-range proteinuria at time-zero biopsy, as well as 1-, 3-, and 5-year renal survival. This rapid, scalable method can be deployable in the form of software at the point of care, and holds the potential for substantial clinical impact, including augmenting clinical decision making for nephrologists. This framework can also be adapted to other organ-specific pathologies focused on evaluating fibrosis, as well as image datasets developed using other histological staining protocols. Further validation of the models across different clinical practices and image datasets is necessary to validate this technique across the full distribution and spectrum of lesions encountered in a typical pathology service.

DISCLOSURE

All the authors declared no competing interests.

ACKNOWLEDGMENTS

This work was supported by grants from the Department of Medicine, Boston University Medical Campus (BUMC), Affinity Research Collaboratives funding (BUMC), Scientist Development Grant (17SDG33670323) from the American Heart Association and Boston University's Hariri Research Award (#2016-10-009) to VBK, Boston University's Undergraduate Research Opportunities Program funding to CL and DM, NIH grants (R01-HL132325 and R01-CA175382) to VCC. VBK thanks Mr. Aaron Freed (BUMC) and Ms. Katia Oleinik (BU) for their assistance on network administration and high performance computing, respectively. The authors thank Dr. Michael Kirber at the BUMC imaging core facility for his assistance.

AUTHOR CONTRIBUTIONS

VBK, JMF, DJS, and VCC conceived and designed the study; PS and DM performed chart review and extracted clinical data; CQL generated digitized images of the kidney biopsies; JMH examined renal biopsy samples and reported fibrosis grading; VBK performed the experiments; VBK and MEB generated the figures; VBK, MEB, JMF, DJS, and VCC interpreted the data and results; VBK and DJS wrote the manuscript; PS, JMF, and VCC edited the manuscript; and DJS and VCC provided overall supervision.

SUPPLEMENTARY MATERIAL

Supplementary Material. Glossary of technical terms, data collection, performance metrics, and data augmentation.

Figure S1. Flowchart representing the project workflow.

Figure S2. Sensitivity analysis on the convolutional neural network (CNN) model.

Figure S3. Sensitivity analyses performed to determine optimal layer for fine-tuning the convolutional neural network (CNN) model.

Figure S4. Performance of the model developed using pathologist-derived estimate of fibrosis.

Figure S5. Predictive model of eGFR at the time of biopsy using $\times$100 magnification images.

Figure S6. Performance of the convolutional neural network (CNN) model is a function of data augmentation.

Table S1. Comparison of convolutional neural network (CNN) and pathologist-estimated fibrosis score (PEFS) models.

Supplementary material is linked to the online version of the paper at www.kireports.org.

REFERENCES

1. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA*. 2016;315:551–552.
2. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920–1930.
3. Kang J, Schwartz R, Flickinger J, Beriwal S. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. *Int J Radiat Oncol Biol Phys*. 2015;93:1127–1135.
4. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216–1219.
5. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. *Am J Gastroenterol*. 2010;105:1224–1226.
6. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal*. 2012;16:933–951.
7. Beck AH, Sangoi AR, Leung S, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*. 2011;3:108–113.
8. Xua J, Luo X, Wang G, et al. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*. 2016;191:214–223.
9. Albarqouni S, Baur C, Achilles F, et al. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans Med Imaging*. 2016;35:1313–1321.
10. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol*. 2017;52:434–440.
11. Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. *Sci Rep*. 2017;7:46450.
12. Han Z, Wei B, Zheng Y, et al. Breast cancer multi-classification from histopathological images with structured deep learning model. *Sci Rep*. 2017;7:4172.
13. Kandaswamy C, Silva LM, Alexandre LA, Santos JM. High-content analysis of breast cancer using single-cell deep transfer learning. *J Biomol Screen*. 2016;21:252–259.
14. Saha M, Chakraborty C, Arun I, et al. An advanced deep learning approach for Ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. *Sci Rep*. 2017;7:3213.
15. Turkki R, Linder N, Kovanen PE, et al. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *J Pathol Inform*. 2016;7:38.
16. Vandenberghe ME, Scott MLJ, Scorer PW, et al. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci Rep*. 2017;7:45938.
17. Wang J, Yang X, Cai H, et al. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci Rep*. 2016;6:27327.
18. Cha KH, Hadjiiski LM, Samala RK, et al. Bladder cancer segmentation in CT for treatment response assessment: application of deep-learning convolution neural network—a pilot study. *Tomography*. 2016;2:421–429.
19. Ciompi F, Chung K, van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep*. 2017;7:46479.
20. Cruz-Roa AA, Arevalo Ovalle JE, Madabhushi A, González Osorio FA. A deep learning architecture for image

representation, visual interpretability and automated basal-cell carcinoma cancer detection. *Med Image Comput Comput Assist Interv*. 2013;16:403–410.

21. Danaee P, Ghaeini R, Hendrix DA. A deep learning approach for cancer detection and relevant gene identification. *Pac Symp Biocomput*. 2016;22:219–229.
22. Inglese P, McKenzie JS, Mroz A, et al. Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chem Sci*. 2017;8:3500–3511.
23. Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans Comput Biol Bioinform*. 2015;12:928–937.
24. Qiu J, Yoon HJ, Feam PA, Tourassi GD. Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE J Biomed Health Inform.*, in press.
25. Sirinukunwattana K, Raza SEA, Tsang YW, et al. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans Med Imaging*. 2016;35:1196–1206.
26. Sun W, Zheng B, Qian W. Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis. *Comput Biol Med.*, in press.
27. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci Rep*. 2017;7:5301.
28. Yuan Y, Shi Y, Li C, et al. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics*. 2016;17(suppl 17):476.
29. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
30. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Available at: https://arxiv.org/abs/1512.00567. Accessed October 5, 2017.
31. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
32. Zhang CY, Zhou P, Li C, Liu L. A convolutional neural network for leaves recognition using data augmentation. Cit/Iucc/Dasc/Picom 2015 IEEE International Conference on Computer and Information Technology—Ubiquitous Computing and Communications—Dependable, Autonomic and Secure Computing—Pervasive Intelligence and Computing; 2015:2147–2154.
33. Simard PY, Steinkraus D, Platt JC. Best practices for convolutional neural networks applied to visual document analysis. Proceedings of the Seventh International Conference on Document Analysis and Recognition, Vols. I and II. 2003:958–962.
34. Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process Lett*. 2017;24:279–283.
35. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60:84–90.
36. Ding J, Chen B, Liu H, Huang M, et al. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geosci Remote Sensing Lett*. 2016;13:364–368.
37. Cui XD, Goel V, Kingsbury B. Data augmentation for Deep Convolutional Neural Network Acoustic Modeling. 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2015:4545–4549.
38. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975;405:442–451.
39. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20:37–46.
40. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85:257–268.
41. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. 2005;37:360–363.
42. Becker G. Creating comparability among reliability coefficients: the case of Cronbach alpha and Cohen kappa. *Psychol Rep*. 2000;87:1171–1182.
43. Mariani LH, Martini S, Barisoni L, et al. Interstitial fibrosis scored on whole-slide digital imaging of kidney biopsies is a predictor of outcome in proteinuric glomerulopathies. *Nephrol Dial Transplant.*, in press.
44. D'Amico G, Ferrario F, Rastaldi MP. Tubulointerstitial damage in glomerular diseases: its role in the progression of renal damage. *Am J Kidney Dis*. 1995;26:124–132.
45. Marcussen N. Tubulointerstitial damage leads to atubular glomeruli: significance and possible role in progression. *Nephrol Dial Transplant*. 2000;15(suppl 6):74–75.
46. Nath KA. Tubulointerstitial changes as a major determinant in the progression of renal damage. *Am J Kidney Dis*. 1992;20:1–17.
47. Rodríguez-Iturbe B, Johnson RR, Herrera-Acosta J. Tubulointerstitial damage and progression of renal failure. *Kidney Int*. 2005;(Suppl 99):S82–S86.
48. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016;7:29.
49. Romo-Bucheli D, Janowczyk A, Gilmore H, et al. A deep learning based strategy for identifying and associating mitotic activity with gene expression derived risk categories in estrogen receptor positive breast cancers. *Cytometry A.*, in press.
50. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol.*, in press.
51. Street JM, Souza AC, Alvarez-Prats A, et al. Automated quantification of renal fibrosis with Sirius Red and polarization contrast microscopy. *Physiol Rep*. 2014;2, pii: e12088.