National Institute of Environmental Health Sciences
Division of Translational Toxicology

Assessment and Validation of Deep Learning Algorithms in Identifying Early Chronic Progressive Nephropathy Changes

Priyanka Thakur[1,2], David Cunefare[1,2], Charan Ganta[1,3], Cynthia Willson[1,3], Allison C. Boone[1,4], Katherine Allen-Moyer[5], Keith Shockley[6], Eli Ney[1], Ronald Herbert[1], Mark Cesta[1]

[1]Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

[2]Charles River Laboratories, Durham, NC, USA, [3]Inotiv, Morrisville, NC, USA, [4]Experimental Pathology Laboratories, Inc., Morrisville, NC, USA, [5] Social and Scientific Systems, Inc., a DLH Holdings Corp Company, Durham, NC, USA, [6]Division of Intramural Research, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Abstract

Introduction: Convolutional neural networks (CNNs) based deep learning (DL) methods of artificial intelligence (AI) have rarely been used for lesion identification in diagnostic histopathology and have the potential to identify early changes of chronic progressive nephropathy (CPN) in rodent studies. **Methods and Materials:** Hematoxylin and eosin (H&E) stained kidney tissue sections from two sub-chronic rodent Division of Translational Toxicology (DTT) studies were retrospectively evaluated for early CPN changes using AI. **Experimental Design**: The initial CPN AI algorithm (APP1) was trained on regions of interest (ROIs) extracted from 23 whole slide images (WSIs) from a single study (Study 1). In a second AI algorithm (APP2), we modified our original algorithm by integrating extra training slides from another study (study 2). Two board-certified veterinary pathologists annotated validation data sets using the diagnostic criteria established for training. Dice coefficient, sensitivity, and precision metrics were calculated to assess the agreement between the AI and pathologist annotations. **Results**: Compared to APP1, APP2 increases precision for study 2. However, other metrics (dice coefficient and sensitivity) were equivalent between APP1 and APP2 across both studies. Both algorithms detected a greater number of CPN lesions than the pathologists. **Conclusion:** Compared to APP1, APP2 reduced the over-detection of not true CPN lesions in study 2 and increased precision, which indicates that five additional training slides from the study can impact the algorithm's rate of over-detection within that study. **Impact Statement**: These algorithms could increase the sensitivity for detecting early CPN and reduce the time spent by pathologists in identifying subtle early CPN lesions to provide quick decision support.

Introduction

Early CPN changes, often subtle, may exist as spontaneous background lesions or be test article-related. This study used AI to significantly alleviate the pathologist's laborious task of identifying these subtle background CPN lesions and compared performance of two different algorithms: one based on a single study (APP1) and one that also included a second study (APP2).

The Dice Coefficient, sensitivity, and precision metrics demonstrate the efficiency of AI algorithms in detecting CPN lesions. They provide a measure of how accurately the pathologists' CPN annotations align with the algorithms' CPN detections, thereby offering a comprehensive measure of the AI algorithm's performance.

We also quantified the areas where AI-identified CPN lesions overlapped with pathologist annotations, regardless of the accuracy of the overlap. Additionally, we identified CPN lesions that were detected solely by AI algorithms but not annotated by pathologists, further highlighting the potential of AI in this field.

We calculated whole lesion performance metrics:

1.) The number of pathologists' CPN annotations overlapping with AI detections

2.) Additional CPN detections by the algorithms (APPs) that pathologists did not mark

3.) Pathologist only detections that were missed by the AI algorithms (APPs)

These metrics capture the agreement between the algorithms and the pathologists for marking the same locations for lesions, without factoring in the accuracy of the lesion boundaries which can often be difficult to define.

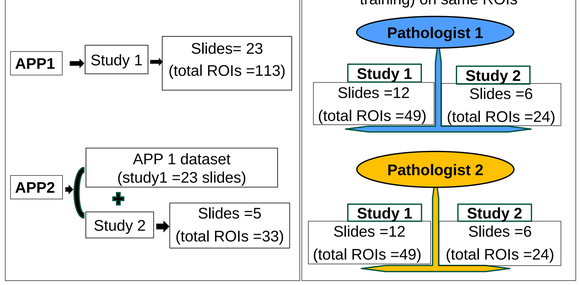
Methods

Image analysis was done using an Image Analysis Software platform. Both APP1 and APP2 used DeepLab CNNs at 20x magnification to segment the lesion areas and were trained using the same parameters. APP1 was trained on annotations in 23 WSIs, and APP2 was trained on those same 23 WSIs and an additional 5 from another study.

After training the APPS, they were run on 18 validation images that did not overlap with the training set. The results of the APPs were compared to the expert manual annotations to calculate Dice Coefficient, sensitivity, and precision metrics.

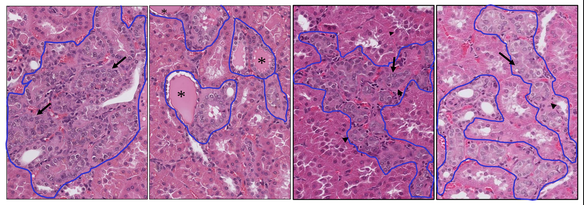
Experimental Design

Training	Validation
(Training annotations done by two pathologists)	Validation dataset annotations done by two pathologists (not involved in training) on same ROIs



Diagnostic criteria for early CPN changes

1. Focal to multifocal foci of tubular basophilia with or without simple tubule hyperplasia
2. Evidence of tubular regeneration in outer kidney (nuclear crowding/karyomegaly with cytoplasmic basophilia)
3. Peritubular basement membrane thickening
4. Protein/hyaline casts, often first observed in outer medulla
5. Variable infiltration by mononuclear inflammatory cells associated with tubular changes



Representative examples of training annotations for CPN lesions (indicated by blue markings)
- Nuclear crowding and tubular basophilia (arrow)
- Protein/hyaline casts (asterisk)
- Peritubular basement membrane thickening (arrowhead)

Fig. 1: Dice coefficient, Sensitivity and Precision

Precision = Area of overlap (TP) / overlap (TP) + FP (additional detections by algorithm)

Sensitivity = Area of overlap(TP) / overlap (TP) + FN (additional detections by pathologist)

Dice Coefficient* = 2 TP / (2 TP + FP + FN) (F1 score)

As pathologists performed similarly across metrics, data from the two pathologists were pooled for analysis

**The Dice coefficient, also known as the Dice Similarity Coefficient, is used to measure the similarity between two sets of data, such as segmentations of an image. It ranges from 0 to 1, with 1 indicating a perfect match and 0 indicating no overlap.*

Precision

Row Labels	Average	Std.Dev.	Min	Max	Median
APP 1	**0.711**	**0.113**	**0.355**	**0.917**	**0.690**
Study 1	0.704	0.123	0.355	0.917	0.685
Study 2	0.725	0.091	0.569	0.854	0.722
APP 2	**0.782**	**0.116**	**0.413**	**0.950**	**0.813**
Study 1	0.738	0.117	0.413	0.928	0.709
Study 2	0.870	0.044	0.785	0.950	0.866
Grand Total	**0.747**	**0.119**	**0.355**	**0.950**	**0.739**

Sensitivity

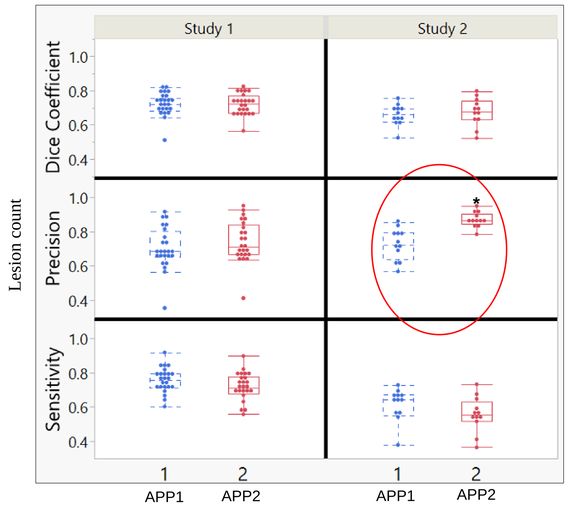
Row Labels	Average	Std.Dev.	Min	Max	Median
APP 1	**0.707**	**0.102**	**0.379**	**0.917**	**0.717**
Study 1	0.753	0.070	0.603	0.917	0.757
Study 2	0.615	0.094	0.379	0.728	0.645
APP 2	**0.663**	**0.115**	**0.366**	**0.897**	**0.687**
Study 1	0.716	0.080	0.558	0.897	0.711
Study 2	0.556	0.102	0.366	0.733	0.554
Grand Total	**0.685**	**0.110**	**0.366**	**0.917**	**0.701**

Dice

Row Labels	Average	Std.Dev.	Min	Max	Median
APP 1	**0.697**	**0.070**	**0.512**	**0.818**	**0.698**
Study 1	0.718	0.067	0.512	0.818	0.720
Study 2	0.656	0.059	0.525	0.755	0.659
APP 2	**0.702**	**0.069**	**0.524**	**0.814**	**0.696**
Study 1	0.717	0.059	0.566	0.814	0.722
Study 2	0.673	0.080	0.524	0.793	0.675
Grand Total	**0.700**	**0.069**	**0.512**	**0.818**	**0.698**

Fig. 2: Precision

Compared to APP1, APP2 increases precision for study 2 slides

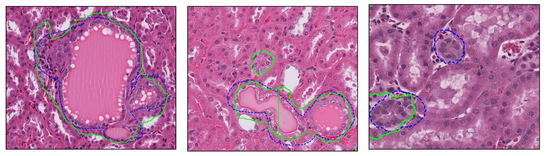


**t-test was performed to compare algorithm performance across these metrics for both studies. The precision metric within study 2 was significant (p-value =0.0002)*

Fig. 3: Representative examples area overlap and non-

For the whole lesion analysis, any manual annotation that had an overlap of at least 300 μm² with an automatically detected lesion was considered detected by the algorithm, and the rest were considered missed. Any automatically detected lesion that did not overlap at least 300 μm² with a manual annotation was considered an extra detection by AI algorithm.

Representative examples of overlapping CPN detections, Algorithm's detection and pathologist only detections are shown in images below:



Overlapping CPN detections (Blue dotted line: pathologist's Annotations (Green: AI detections)

Algorithm's only CPN detections (non-overlapping green)

Pathologists' only annotations (non-overlapping blue dotted line)

Fig. 4: Overlap, pathologist-only, and AI-only detections

- In general, there was good overlap between annotations made by the pathologists and by the algorithms (APP1 &APP2) for both study 1 and study 2
- Some CPN lesions were only detected by pathologists
- APP1 and APP2 also detected subtle CPN lesions which were not annotated by pathologists

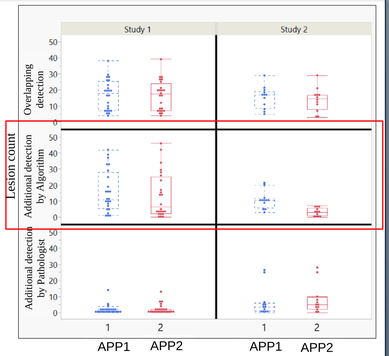


Fig. 5: Some AI-only detected regions were not CPN lesions or borderline

A third board-certified pathologist further analyzed AI (APP1 & APP2) detected CPN lesions (not annotated by pathologists) to determine how many are actual CPN lesions and how many are not. This analysis categorized AI-additional detections into three classes:

- **True CPN**: additional detections that meet our diagnostic criteria
- **Not CPN**: additional detections which are not CPN (e.g., blood vessels, extra eosinophilic vascular fluid, a collection of hypertrophied basophilic cells with unclear histological structure, inflammatory cell infiltrate with no associated tubular change)
- **Borderline:** The number of additional detections that do not meet the diagnostic criteria for identifying CPN such as cases where only 2-3 renal tubular epithelial cells have basophilic cytoplasm while the rest are normal

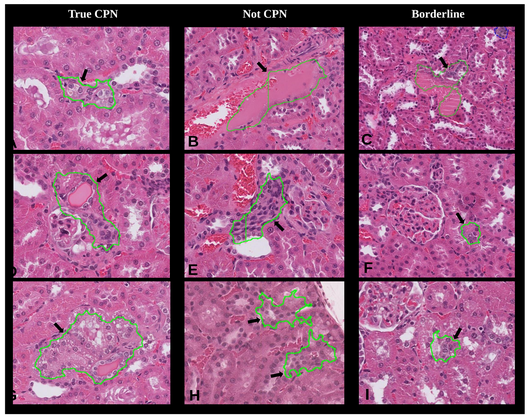
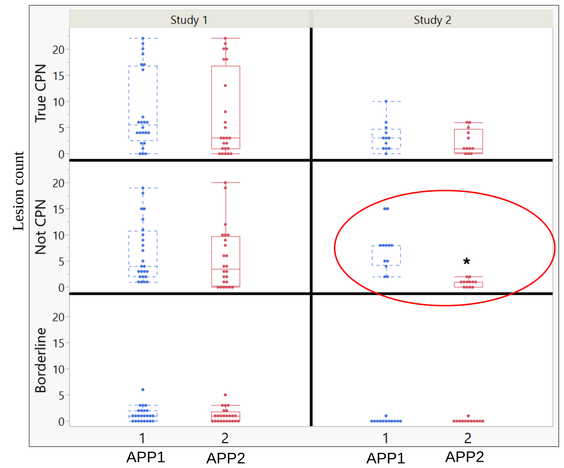


Fig. 6 APP2 outperformed APP1 in detecting "true" CPN

When the algorithms (APP1 and APP2) make a detection that the pathologists do not, APP2 detects fewer non-CPN lesions for study 2 than APP1 while still detecting the same amount of CPN lesions (true CPN and borderline CPN)



**Kruskal–Wallis test was used to assess differences between algorithms across study. The amount of non-CPN lesions detected was different across algorithms (p < 0.0001*

Summary

- Both APP1 and APP2 detected subtle CPN lesions which are not annotated by pathologist
- Both the APPs also detected borderline CPN lesions which were excluded from diagnostic criteria e.g., a few basophilic cells lining the normal renal cortical tubules (Figure 5 C, F & I)
- Both the APPs, however, also over-detected CPN lesions which are not true CPN lesions e.g., sections of blood vessels, extra eosinophilic vascular fluid and normal deeply stained cortical tubules from study 2 (Figure 5 H)
- Compared to APP1, APP2 reduced the over-detection of not true CPN lesions in study 2 (Figure 5 H) and increased precision, which indicates that five additional training slides from the study can impact the algorithm's rate of over-detection within that study
- All other metrics (sensitivity and dice coefficient) were equivalent between APP1 and APP2
- Among the additional detections, APP2 is more precise than APP1, as verified by a third board-certified pathologist
- Whole lesion analysis allowed us to identify the areas where both AI and pathologists excelled in CPN detection and the areas where they fell short
- Future Directions: AI can not only assist pathologists in detecting subtle, early lesions that are easy to miss and time-consuming to identify but also has the potential to subcategorize them

Acknowledgments

The authors thank Ms. Eli Ney and Ashley Paragone of the Comparative and Molecular Pathogenesis Branch Digital Imaging Core and the staff of the NTP archive for their invaluable contribution in providing whole slide image scans of the kidney lesions. The authors also thank the statistical work supported by the National Institute of Environmental Health Sciences under contract GS-00F-173CA / 75N96022F00055 to Social and Scientific Systems, Inc., A DLH Holdings Corp Company.

References

- Meyer J, Khademi A, Tetu B, Han W, Nippak P, Remisch D. Impact of artificial intelligence on pathologists' decisions: an experiment. J Am Med Inform Assoc. 2022 Sep 12;29(10):1688-1695. doi: 10.1093/jamia/ocac103. PMID: 35751441; PMCID: PMC9471707
- Frazier KS, Seely JC, Hard GC, Betton G, Burnett R, Nakatsuji S, Nishikawa A, Durchfeld-Meyer B, Bube A. Proliferative and nonproliferative lesions of the rat and mouse urinary system. Toxicol Pathol. 2012 Jun;40(4 Suppl):14S-86S. doi: 10.1177/0192623312438736. PMID: 22637735
- Hard GC, Banton MI, Bretzlaff RS, Dekant W, Fowles JR, Mallett AK, McGregor DB, Roberts KM, Sielken RL Jr, Valdez-Flores C, Cohen SM. Consideration of rat chronic progressive nephropathy in regulatory evaluations for carcinogenicity. Toxicol Sci. 2013 Apr;132(2):268-75. doi: 10.1093/toxsci/kfs305. Epub 2012 Oct 26. Erratum in: Toxicol Sci. 2013 Jun;133(2):344. PMID: 23104430; PMCID: PMC3595520