Quantifying acute kidney injury in an Ischaemia-Reperfusion Injury mouse model using Deep Learning-based semantic segmentation in histology

Andreea Luchian[1,*], Katherine Trivino Cepeda[2,3], Rachel Harwood[4], Patricia Murray[2,3], Bettina Wilm[2,3], Simon Kenny[4], Paola Pregel[5,] Lorenzo Ressel[1]

[1]Department of Veterinary Anatomy Physiology and Pathology, Institute of Infection, Veterinary and Ecological Sciences, Faculty of Health & Life Sciences, University of Liverpool
[2]Department of Molecular Physiology and Cell Signalling, Institute of Systems, Molecular and Integrative Biology, University of Liverpool
[3]Centre for Pre-clinical Imaging, Institute of Systems, Molecular and Integrative Biology, University of Liverpool
[4]Department of Paediatric Surgery, Alder Hey in the Park
[5]Department of Veterinary Sciences, University of Turin

[*]Corresponding author: Andreea Luchian (Master of Science, MSc), ORCID: 0000-0001-7959-6190, Email: A.Luchian@liverpool.ac.uk, Phone number: +447495658616
Department of Veterinary Anatomy Physiology and Pathology, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Chester High Road, Neston, CH64 7TE

Keywords: “deep-learning”, “ischaemia-reperfusion injury”, “mouse” and “kidney”

Summary statement

This study employs AI-driven analysis of kidney tissue slides, enhancing the understanding of disease progression using mouse models. Deep learning's potential to improve injury assessment methods is explored, challenging conventional standards.

Abstract

This study focuses on Ischaemia-Reperfusion Injury (IRI) in kidneys, a cause of acute kidney injury (AKI) and end-stage kidney disease (ESKD). Traditional kidney damage assessment methods are semi-quantitative and subjective. This study aims to use a Convolutional Neural Network (CNN) to segment murine kidney structures after IRI, quantify damage via CNN-generated pathological measurements, and compare this to conventional

scoring. The CNN was able to accurately segment the different pathological classes, such as Intratubular Casts and Tubular Necrosis, with an F1 score of over 0.75. Some classes, such as Glomeruli and Proximal Tubules, had even higher statistical values with F1 scores over 0.90. The scoring generated based on the segmentation approach statistically correlated with the semiquantitative assessment (Spearman Correlation coefficient = 0.94). The heatmap approach localised the intratubular necrosis mainly in the outer stripe of the outer medulla, while the tubular casts were also present in more superficial or deeper portions of the cortex and medullary areas. This study presents a CNN model capable of segmenting multiple classes of interest, including acute IRI-specific pathological changes, in a whole mouse kidney section and can provide insights into the distribution of pathological classes within the whole mouse kidney section.

Introduction

Acute kidney injury (AKI) is a global public health problem, with rising incidence and mortality rates in recent decades (Lameire et al., 2013). AKI is caused by various pathological processes, including ischemia-reperfusion injury (IRI), follow-on effects of kidney transplantation, drug toxicity, sepsis, and other insults (Lameire et al., 2013). AKI resulting from IRI has complex pathogenesis, and experimental animal models of kidney IRI offer the possibility to examine key pathogenesis-related morphological changes. The gold standard to assess kidney injury remains semi-quantitative histopathology scoring; however, traditional scoring systems are time-consuming, in some cases subject to inter-observer variability and are based only on a small sample area of the kidney section (Wang et al., 2005), hence lacking quantitative power. In addition, kidney tissue has a complex non-redundant architecture, rendering semi-quantitative scoring intrinsically challenging due to the potential lack of representativeness of randomly selected areas of interest. Moreover, the scoring systems available differ in terms of morphological structures and areas analysed (Hesketh et al., 2014; Wang et al., 2005), and a comparative analysis of their performance is lacking.

Digital image analysis techniques can improve the visual assessment of a wide range of images (Kshirsagar and Joshi, n.d.), including histological microphotographs (Niazi et al., 2019). The advantages of using computer algorithms include increased reproducibility and, with sufficient computational power, the ability to analyse the whole sample or multiple batches of samples in automation. As visual assessment is monotonous and prone to lack of objectivity, it represents a significant limiting factor in extensive studies, and the use of automated assessment could overcome this. The traditional approach to performing image analysis refers to computer vision methods such as feature descriptors for object detection (O'Mahony et al., 2020). For tasks such as image classification, a feature extraction step is needed. The main inconvenience with this method is that it is necessary to choose "*a-priori*", which features are valuable in each image. Therefore, as the number of classes increases, this step becomes burdensome. These methods have been successfully used in small kidney histopathology studies; one study developed a segmental histogram of oriented gradients that successfully performed a comprehensive detection of glomeruli in whole kidney sections (Kato et al., 2015) while another study (Grimm et al., 2003) used computerised image analysis of Picro Sirius Red stained kidney tissue sections to quantify the extent of collagen and consequently interstitial fibrosis, a valuable predictor of long-term graft function. Although these "traditional" image analysis techniques can be useful to answer specific research

questions, it would be challenging to automatically apply them in large-scale studies with data sets that are likely to present variations. Moreover, it is up to the computer vision scientist to decide which features best define the various classes of objects after a time-consuming trial-and-error approach. On top of that, a plethora of parameters are required to define each feature, all of them having to be fine-tuned by the operator (O'Mahony et al., 2020). Finally, in some cases, traditional machine-learning approaches may not detect complex morphological features.

In the last two decades, digital imaging has seen the emergence and progress of whole slide imagining (WSI), which permits full slides to be digitised and stored at high resolution (Niazi et al., 2019). More recently, with the advent of the graphics processing unit (GPU)-based computation and Convolutional Neural Networks (CNNs), a full histological section can be completely and consistently analysed for several objects of interest, using a supervised training strategy without the need to pre-set the features for each object. CNNs represent a deep learning method that draws inspiration from the intricate organisation of human brains. By employing a model structure comprising multiple processing layers, deep learning allows for the acquisition of diverse levels of data representation, leading to unprecedented improvements in model performance. This state-of-the-art technology has revolutionised various fields, including speech recognition, visual object identification, and drug discovery and genomics domains (Xie et al., 2020). CNNs are the most applied deep-learning models for bio-image analysis. CNNs mimic the human visual perception process via a cascade of interconnected, layered units (neurons) that resemble the visual system architecture (Lindsay, 2021; Tang et al., 2019). In contrast with the traditional approaches for image analysis, neural networks can be trained to automatically detect underlying patterns in classes of images that have been previously labelled and extract and the most descriptive features (O'Mahony et al., 2020). Previous work involving CNNs applied to WSIs focused on classifying human cancers, such as invasive breast cancer (Cruz-Roa et al., 2017) and identifying pancreatic endocrine tumours (Niazi et al., 2018), among others. In animal models, CNNs have also been applied to histologically score lung fibrosis and inflammation(Heinemann et al., 2018) and to assess the severity of various pulmonary lesions (Asay et al., 2020). The utilisation of CNNs in the domain of kidney histopathology is a relatively recent development, and the existing body of literature primarily focuses on the detection of glomeruli (Bukowy et al., 2018; Gallego et al., 2018). The latest investigations pertaining to human kidney biopsies have also directed their attention towards the segmentation of multiple classes of interest within distinct scenarios. For instance, some studies have specifically targeted the segmentation of classes within healthy kidney tissue (Marechal et al., 2022), while others have concentrated on biopsies associated with IgA-nephropathy (Hölscher et al., 2023) and chronic kidney injury (Ginley et al., 2021). Furthermore, certain studies have exclusively addressed the segmentation of the healthy kidney cortex or both the cortex and medulla, omitting the inclusion of vital components such as the papilla, transitional epithelium, and associated connective tissue (e.g., stroma and adipose tissue), which frequently appear in the observed sections (Bukowy et al., 2018; Hermsen et al., 2019). However, no study so far has included the CNN-based segmentation on a specific disease model quantifying both injured and healthy structures on an entire mouse kidney section. The present study aimed at applying CNNs to kidney WSI and testing their efficacy to detect levels of kidney damage in an IRI mouse model using a segmentation approach and to compare it to a widely used traditional semi-quantitative method.

Results

CNN Multiclass Segmentation Performance

Multiclass semantic segmentation of kidney sections enabled the extraction of quantitative histological features on a large scale. An example of a fully segmented pathological mouse kidney section is depicted in Fig. 1 (a healthy segmented kidney is presented in Fig. S1).

The segmentation performance on the test set was assessed via a confusion matrix (Fig. 2) from where precision, recall, specificity and F1 were extracted (Table 1). Good performances were obtained for the “Glomeruli” class, where 94% of the ground truth labels were correctly identified with a precision of 0.99. Segmentation performances for “Stroma”, “Transitional epithelium”, and “Adipose tissue” were similar to the “Glomeruli” class (above 90%) but with lower precision. For the "Proximal Tubule" class, 88% of all pixels labelled with this name were correctly classified by CNN with a precision >0.95. As regards the classes representing pathological changes, "Intratubular casts" and "Tubular necrosis", 70% and 85% of the ground truth pixels were correctly classified, respectively, with a precision of 0.87 and 0.94. The overall model statistics are presented in Table S3.

Lower precision values were observed for the "Regenerating epithelium" class, and the misclassification was mainly with the "Distal tubules/Collecting ducts" class (44%). The "Regenerating epithelium" class was further excluded from the generation of the CNN-based scoring due to its lower statistical values.

CNN-based IRI scoring vs semi-quantitative IRI scoring method and heatmap

Two classes (“Intratubular Casts” and “Tubular Necrosis”) were selected as representative of pathological changes as they presented good statistical values (F1=0.78 for “Intratubular Casts” class and F1=0.89 for “Tubular Necrosis” class). The network applicability to score IRI damage was assessed by comparing CNN's quantification of selected classes, "Tubular Necrosis" and "Intratubular casts" (Table S2), to the semi-quantitative scoring method. Upon heatmap analysis (Fig. 3), the classes “Intratubular casts” and “Tubular Necrosis” were clearly spatially identified. Areas of more intense necrosis were often focused in between the cortex and medulla (Outer stripe of outer medulla; OSOM), a region known to be affected by hypoxic damage (2,3). Casts were often associated with the same area but varied in localisation, involving more superficial or deeper portions of the cortex and medullary areas.

The conventional semiquantitative method of scoring appeared highly correlated to the CNN scoring with a Spearman r (31) =0.94 ($p<0.0001$) (Fig. 4).

Discussion

Acute IRI models are, at present, evaluated morphologically using semi-quantitative histological methods. Limitations of this approach include lack of reproducibility, interobserver variability and limited sample analysis. Additionally, the systems used may differ in terms of the morphological structures and areas that are analysed (Hesketh et al., 2014; Sun et al., 2016; Wang et al., 2005). Here, we have developed a novel system to quantify kidney damage in a mouse model of IRI by using a DL approach. Compared to the traditional way of scoring IRI, where only ten fields of view focused on the OSOM are being used, which represents approximately 10% of the kidney surface, our methods use WSIs that allow a full assessment of the mouse kidney section, increasing the quantitative power of the scoring method. We have obtained good segmentation results for the majority of classes of interest, with 9/10 presenting an F1 value higher than 0.70.

With an overall model precision of 0.85, these results align with previous work on human kidneys (Hermsen et al., 2019). The present study represents the first deep learning model to simultaneously segment and classify nine classes of interest on a full mouse kidney section originating from ischaemia-reperfusion injury surgery. Previous work focused only on detecting glomeruli (Gadermayr et al., 2019) in mouse sections or creating multiple networks for different morphological structures in human samples (Jayapandian et al., 2021), possibly increasing the time required to segment and classify all structures. Using a single network to identify multiple structures accurately decreases the overall time needed for full classification. The DL model described here was trained and tested on single stain (PAS) WSIs, compared to Jayapandian and colleagues (Jayapandian et al., 2021), who tested a CNN approach on multiple stains (Hematoxylin & Eosin, PAS, Silver, and Trichrome). This also represents an advantage of our system in terms of time and material needed for the analysis. The selection of PAS staining was based on its ability to provide a more precise evaluation of the basement membrane when compared to H&E staining. Furthermore, PAS staining was preferred due to the increased positivity exhibited by the proteinaceous casts in contrast to H&E staining, where the casts appear significantly lighter (Dvanajscak et al., 2020). This study utilised pathological sections from three IRI experiments, with the experimental design being consistent with one another but conducted at different time points. A fourth experiment had a distinct experimental setup, featuring unilateral rather than bilateral ischemia and a longer duration. This study used only the unclamped kidneys from the unilateral IRI experiment as healthy controls. Additionally, sections were processed in two separate laboratories to account for domain shift, encompassing variations in experimental conditions and processing techniques such as embedding. It is important to note that the PAS staining was performed in the same laboratory, thereby limiting the debate on staining variability in this study. However, colour normalisation can be utilised in cases where staining intensity varies. Moreover, the DL model performed well on both healthy and pathological sections. The best segmentation performance was achieved for the "Glomeruli" class (F1>0.95), as similarly achieved in previous studies on the human kidney (Hermsen et al., 2019).

The DL model also demonstrated great potential to distinguish between proximal (F1>0.90) and "Distal tubules/Collecting ducts" (F1>0.70). Although the focus of the work was to achieve a system to quantify pathological classes, the identification of multiple healthy classes compared to a single "healthy kidney tissue" class opens the option of more detailed types of analyses, such as the quantification of the amount of damage per number of glomeruli, per nephron or tubular area. In addition, this approach allows spatial analysis (e.g., the

average distance between specific healthy areas and pathological ones) between different classes using a heatmap approach which can complement the quantitative analysis of the whole section.

In histopathology, overfitting can occur when the DL model has learned to recognise the unique features of the training images set, but it does not generalise well to new images that contain different variations of the same tissue structure. In generating the CNN injury score, we deemed it appropriate to extract quantification results from both our training and testing datasets, as overfitting does not represent a crucial issue when the task is to accurately quantify an experimental dataset of WSIs from which annotation examples will be generated. In this regard, each WSI offers training and testing areas within the same slide. Two pathological classes, namely "Intratubular casts" and "Tubular necrosis", were considered reliable and were further used to score damage in the mouse kidney sections. These two classes represent the most widely used parameters in different scoring systems to identify acute IRI changes (Hesketh et al., 2014; Sun et al., 2016; Wang et al., 2005). Necrosis, in particular, overlapped as expected with OSOM in most samples, as mentioned in previous studies (Hesketh et al., 2014; Wang et al., 2005).

"Regenerating epithelium" represented the least successful class to segment efficiently. The regeneration process in the kidney is a complex process encompassing a continuum of pathogenetic transition (degeneration, necrosis, regeneration) and associated different morphological hallmarks (e.g., flattened cells, mitotic figures, and plump cells). Therefore, obtaining the ground truth for this class presented as quite challenging, likely due to the intrinsic variable morphology of the class itself overlapping with a single concept of "regeneration". This translated into minor statistical values (F1>0.20) and would benefit from more training data or a further splint into the morphological stages of this particular change, which is beyond the scope of the present study. In addition, because the mice in the acute IRI model were only kept for three days after surgical induction of IRI, the regenerating process was minimally represented compared to the necrotic phase identified, suggesting that this class should be likely better characterised in subsequent timeframes of the IRI and regeneration process (e.g. sub-acute, chronic) (Frazier et al., 2012). It is important to note that the "Regenerating epithelium" class represented a very small portion of the segmented kidney area (average 0.02%) and was confused mainly with the "Distal tubules/Collecting ducts" class, not interfering significantly with the classes we used to score IRI damage. Pitfalls related to morphologically difficult classes are expected in this type of study, as seems to be the case in the work of Hermsen and colleagues (Hermsen et al., 2019), where the class "empty bowman capsule" is likely overlapping large veins according to the segmentation masks provided.

The CNN-based scoring was compared to a previously used semiquantitative way of scoring performed by a board-certified pathologist. The results support a positive correlation between the pathologist and the DL model for detecting IRI severity in an acute mouse model, suggesting that the algorithm correctly assesses general features that are accepted signs of IRI by pathologists. As a general trend, the DL algorithm assigned lower grades than the pathologist. This could be because the area analysed has been dramatically extended, and/or the traditional way of scoring looks only at the OSOM, a region known to be affected by hypoxia (Hesketh et al., 2014; Wang et al., 2005), ignoring all the other kidney regions. In addition, it is important to consider that the pathologist's personal perception may unintentionally lead to the selection of more severely injured areas for analysis. However, the pathologist and the CNN scorings overlap in the healthy non-injured kidney sections

(n=3) as both assigned a score of 0. It is important to note that the inter-observer variability issue was not addressed in this study, as the quantification by the CNN was only compared to the opinion of one pathologist.

Heatmaps were drawn based on the quantification of “Intratubular casts” and “Tubular necrosis” of each grid patch. As expected, most of the damage was localised in the OSOM region of the kidney (Hesketh et al., 2014; Sun et al., 2016; Wang et al., 2005). However, in several cases, a substantial amount of damage (mainly cast formation) was visually identified in cortical and medullary areas, indicating that traditional scoring systems that analyse fields of view only from the cortex or only from the OSOM might not provide the most accurate assessment.

Many studies developed DL models to detect and grade tumours in the past decade. In kidney histopathology, the morphological changes associated with disease (e.g., IRI, allograft rejection) are more complex in appearance compared to the more spatially homogeneous morphological landmarks of the neoplastic process; from this point of view, our approach represents a useful DL algorithm for the segmentation and classification of renal structures that is applicable to the preclinical field of IRI.

This study presents a CNN model capable of segmenting and classifying multiple classes of interest, including acute IRI-specific pathological changes, in a whole mouse kidney section. Moreover, the DL model was applied to sections from different IRI experiments, suggesting that the model generalises well and can represent a useful tool for quantitatively investigating acute IRI models upon histology.

Materials and Methods

Animals and surgery

All experiments were conducted in accordance with the Animals (Scientific Procedures) Act 1986 under a project licence (PPL 7008741 and PP3076489) and were approved by the Animal Welfare and Ethical Review Board (AWERB) of the University of Liverpool. B6 albino mice (C57BL/6J.Tyrc-2J) were purchased from Charles River, Italy, and used to establish a colony that was maintained by the Biomedical Services Unit (BSU) at the University of Liverpool, UK. Mice were housed in ventilated cages with a 12-hour light/dark cycle and access to water and food ad libitum.

The mouse model of renal IRI was induced by bilateral clamping of the renal pedicle using a dorsal approach. Male mice (9–10 weeks) were anaesthetised (isoflurane 1.5%; 1.0 L/min, O_2) for 30 minutes prior to the surgical procedure (Harwood et al., 2019). The body temperature was controlled at 36.5–37.1 °C with a homeothermic monitor system (PhysioSuite, Kent Scientific, Torrington), and the renal pedicle was carefully dissected and clamped with a non-traumatic vascular clamp (InterFocus Ltd, Liton, 18052-03) for 27.5 minutes. The confirmation of ischaemia was observed through a colour change in the kidney, from red to dark purple. Once the clamp was released, the kidney returned to its normal red colour, indicating reperfusion. Subsequently, the surgical wound was repaired, and the animal allowed to recover in a warmed chamber at 37°C for 30 minutes prior to returning to their cage.

Four different sets of experiments were carried out, and 3 of them followed the procedure described above. The 4th experiment followed the same surgery and anaesthesia methods, except the clamp was unilateral right-sided and for a period of 40 minutes. Only the unclamped kidneys from the unilateral IRI experiment were used from this fourth study as healthy controls.

Animals were sacrificed via cervical dislocation three days post-IRI surgery, kidneys exteriorised after laparotomy and immediately fixed in 10% Neutral Buffered Formalin (Fisher Scientific, Leicestershire, 10463750) for 24 – 72h.

Histology and semi-quantitative scoring

Mice kidneys were fixed in 10% Neutral Buffered Formalin (Fisher Scientific, Leicestershire,10463750) for 24–72h. After fixation, kidneys were sagittally sectioned in two halves and placed in formalin into Slotted Tissue Cassettes (Fisher Scientific, Leicestershire, 15327260) until further processing (Morawietz et al., 2004). The specimens were paraffin-embedded and sectioned, using standard procedures, by the LBIH Biobank (University of Liverpool, Liverpool, UK) or the Veterinary Pathology Laboratory, Department of Veterinary Anatomy and Physiology. Sections of 3-4μm were mounted on glass slides, stained with Periodic Acid-Schiff (PAS) and coverslipped by the Veterinary Pathology Laboratory (Leahurst Campus, University of Liverpool, Wirral, UK) and dried for histological analysis.

A total of 34 mid-coronal renal full sections, originating from 18 animals, were used in the study. These included 28 sections originating from 15 animals (23 clamped kidneys), subject to direct IRI and six sections originating from 3 animals (unclamped kidneys)

Subsequently, the slides were scored using a semi-quantitative scoring as previously performed by the same group (Sharkey et al., 2019), adapting a method previously described by Wang and colleagues (Wang et al., 2005). Briefly, kidney lesions were scored depending on the displayed pathological changes (tubules that displayed typical changes of the IRI model, including cell necrosis, intratubular cast formation, reduction of brush border, tubular dilation and tubule regeneration) on ten randomly selected fields of view (FOV) from the outer stripe of the outer medulla (OSOM) and cortex. The FOVs were given a score of 0-4 where 0=0%; 1=1-25%; 2=26-59%; 3=51-75%; 4=76%-100%, then the average of 10 FOVs scores were considered the final score for each tissue section. Scorings were performed using a brightfield microscope (Leica Biosystems, Nussloch, Germany) at 200X magnification by a board-certified veterinary pathologist (LR).

Digitalisation and Neural Network training and analysis

PAS-stained slides were digitally scanned using the Aperio CS2 slide scanner (Leica Biosystems, Nussloch, Germany), with Plan Apo 20X objective lens setup, image size ranging from 21000 to 35000-pixel width and 13000-to-31000-pixel height (0.504 microns per pixel), and visualised using ImageScope™ software (Leica Biosystems, Nussloch, Germany).

The WSIs (n=34) were randomly split into training and testing sets as follows: 17 training and 17 testing. The ground truth was created by manually annotating regions corresponding to normal or pathological changes: of The following normal renal structures represented classes as per normal microscopic anatomy: “Background” (area of the slide characterised by the homogeneous white area without the presence of any histological structure), "Adipose tissue” (area of the stroma characterised by large numbers of adipocytes), “Glomeruli” (round structures that represent a complex web of capillaries), “Proximal tubules**”** (elongated structures with abundant, pink cytoplasm and an easily identifiable brush border), "Distal tubules and collecting ducts**”** (tubular structures with wider lumen, no brush border and less pink cytoplasm than proximal tubules in the cortex and tubules within the medulla), **“**Stroma**”** (connective tissue containing fibroblasts) and “Transitional epithelium**”** (multiple cuboidal layers of epithelium within the renal pelvis). Pathological classes selected for the purpose of model training were: “Intratubular casts**”** (uniformly staining proteinaceous material structures found within/filling the tubular lumen), “Tubular necrosis**”** (destruction of tubular epithelial cells shedding into the tubule lumen), “Regenerating epithelium**”** (tubular epithelium with flattened cells, and/or more cuboidal cells and/or mitotic figures). References (Percy and Barthold, 2007; Scudamore, n.d.) were used during annotations of normal and pathological structures. Annotations were performed by one investigator (AL) and reviewed by a board-certified veterinary pathologist (LR). An example of annotated tissue is available in Fig. S1. “Intratubular casts”, “Tubular necrosis”, and “Regenerating epithelium” classes were considered representative of pathological changes associated with IRI. Cutting/staining artefacts were rarely present on the slides and were not included in the annotations. The number of annotations was automatically balanced to a median of 288 per class. The total number of training annotations used to train the CNN model was 2880 in total (Table S1).

The deep learning process took approximately eight days on a system equipped with 4x Nvidia® Quadro® RTX8000 GPUs (Nvidia, Santa Clara, California) using dedicated software MIMPro (Medical Image Manager Pro with Deep Learning Add On; HeteroGenius®). The employed CNN model is composed of the descending, downsampling portion of a UNET architecture(Ronneberger et al., 2015), where upsampling layers were not included with an 8x downsampled mask as an output. The network was trained for a total of 2300 epochs on the training set, at one iteration per epoch, with batch sizes of 1, 32 and 64. Patches of 512x1024 pixels and magnification of 20X (18167 X 17001 pixels) were used. Adaptive learning rates between 5e-7 and 5e-4 and momentum with values of 0.9 and 0.99 were used depending on monitoring error curve progress "on the fly". Data balancing, dropout (ranging from 0 to 0.1) and random transforms were utilised to increase generalisation and improve the algorithm's robustness for variation in tissue section morphology and staining intensity.

The model obtained was deployed through MIMPro® to create a segmentation where individual pixels are assigned to one of the pre-defined classes. Subsequently, an overlay image (mask) was created where each classified pixel was assigned a colour relating to its class, making the quantification of the pre-defined classes possible. The model was then tested on the test set.

Application of CNN to score acute IRI lesions

All classified areas in pixel except "Background" were summed together for each WSI, representing the surface of the kidney section. The number of pixels representing the pathological classes was transformed into percentages as follows:

$$(Number\ of\ pixels\ classified\ as\ "sum\ of\ pathological\ classes"\ class) /(number\ of\ pixels\ representing\ the\ surface\ of\ the\ kidney\ section)\ x100 = Percentage\ of\ pixels\ classified\ as\ "pathological\ classes"\ .$$

The Percentage of pixels classified as "pathological classes" was considered the score assigned by the CNN model.

To visually map the quantified pathological areas of each single pathological class within the processed WSI, a grid of patches was created and overlayed on the digitally scanned slides (MIMPro®). For each patch of 512X512 pixels, the area covered by the pathological classes was quantified using the previously developed CNN and expressed as a total number of pixels belonging to the class of interest per patch. Pathological classes were then spatially visualised within the kidney parenchyma using a heat map where in a gradient, low values of the classes of interest are represented in blue and high values in red.

Data analysis

To summarise the algorithm's performance, a multiclass confusion matrix was created, from where precision (fraction of predictions as true positives), recall (sensitivity), specificity and F1 (harmonic mean between precision and recall) values were extracted. After the true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) were estimated using a confusion matrix, precision, recall and specificity were calculated using the following formulas: $Precision = TP/(TP + FP);\ Recall = TP/(TP + FN)]; Specificity = TN/(TN + FP)$.

The F1 was calculated using the following: $2 \times PrecisionxRecall/(Precision + Recall) = 2TP/(2TP + FP + FN)$

Pearson's correlation coefficients were calculated among the traditional Wang and CNN-based scoring on both datasets (training and testing). Significance was set as $P<0.05$.

Acknowledgements

The authors thank NVIDIA for GPU support, Derek Magee for software assistance, Charles Milford for technical support in the DiMo laboratory, the University of Liverpool – Leahurst histology laboratory for technical support and Silcock Veterinary Pathology Endowment for supporting slide scanner equipment in the DiMo lab. Alder Hey Children's Kidney Fund supported this work, the European Union's Horizon 2020

Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 813839 and a Kidney Research UK fellowship (TF_010_20171124).

Competing interests

We have no competing interests to disclose.

Funding

This work was supported by Alder Hey Children's Kidney Fund and the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 813839 and a Kidney Research UK fellowship (TF_010_20171124).

Data availability

The deep learning model was created using MIMPro® from HeteroGenius. Short-term licenses can be provided upon request, and the xml file of the model is available at: https://github.com/aluchian/Deep-learning-model.git.

Author contributions statement

Andreea Luchian - literature search, figures, study design, data collection and analysis, data interpretation and writing. Katherine Trivino-Cepeda – planned, executed, and provided material from the bilateral IRI mouse model, writing. Rachel Harwood - performed unilateral IRI surgeries Patricia Murray - supervision, reviewed and edited. Bettina Wilm - supervision, review, and editing. Lorenzo Ressel - Oversight and leadership responsibility for the research activity planning and execution. Simon Kenny – supervision IRI study design, review and editing. Paola Pregel – statistical support, review, and editing.

References

Asay, B.C., Edwards, B.B., Andrews, J., Ramey, M.E., Richard, J.D., Podell, B.K., Gutiérrez, J.F.M., Frank, C.B., Magunda, F., Robertson, G.T., Lyons, M., Ben-Hur, A., Lenaerts, A.J., 2020. Digital Image Analysis of Heterogeneous Tuberculosis Pulmonary Pathology in Non-Clinical Animal Models using Deep Convolutional Neural Networks. Sci Rep 10, 6047. https://doi.org/10.1038/s41598-020-62960-6

Bukowy, J.D., Dayton, A., Cloutier, D., Manis, A.D., Staruschenko, A., Lombard, J.H., Woods, L.C.S., Beard, D.A., Cowley, A.W., 2018. Region-Based Convolutional Neural Nets for Localization of Glomeruli in Trichrome-Stained Whole Kidney Sections. JASN 29, 2081–2088. https://doi.org/10.1681/ASN.2017111210

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N.N.C., Tomaszewski, J., González, F.A., Madabhushi, A., 2017. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Sci Rep 7, 46450. https://doi.org/10.1038/srep46450

Dvanajscak, Z., Cossey, L.N., Larsen, C.P., 2020. A practical approach to the pathology of renal intratubular casts. Seminars in Diagnostic Pathology, Practical Topics and Updates in Renal Pathology 37, 127–134. https://doi.org/10.1053/j.semdp.2020.02.001

Frazier, K.S., Seely, J.C., Hard, G.C., Betton, G., Burnett, R., Nakatsuji, S., Nishikawa, A., Durchfeld-Meyer, B., Bube, A., 2012. Proliferative and Nonproliferative Lesions of the Rat and Mouse Urinary System. Toxicol Pathol 40, 14S-86S. https://doi.org/10.1177/0192623312438736

Gadermayr, M., Dombrowski, A.-K., Klinkhammer, B.M., Boor, P., Merhof, D., 2019. CNN cascades for segmenting sparse objects in gigapixel whole slide images. Computerized Medical Imaging and Graphics 71, 40–48. https://doi.org/10.1016/j.compmedimag.2018.11.002

Gallego, J., Pedraza, A., Lopez, S., Steiner, G., Gonzalez, L., Laurinavicius, A., Bueno, G., 2018. Glomerulus Classification and Detection Based on Convolutional Neural Networks. Journal of Imaging 4, 20. https://doi.org/10.3390/jimaging4010020

Ginley, B., Jen, K.-Y., Han, S.S., Rodrigues, L., Jain, S., Fogo, A.B., Zuckerman, J., Walavalkar, V., Miecznikowski, J.C., Wen, Y., Yen, F., Yun, D., Moon, K.C., Rosenberg, A., Parikh, C., Sarder, P., 2021. Automated Computational Detection of Interstitial Fibrosis, Tubular Atrophy, and Glomerulosclerosis. J Am Soc Nephrol 32, 837–850. https://doi.org/10.1681/ASN.2020050652

Grimm, P.C., Nickerson, P., Gough, J., McKenna, R., Stern, E., Jeffery, J., Rush, D.N., 2003. Computerized Image Analysis of Sirius Red–Stained Renal Allograft Biopsies as a Surrogate Marker to Predict Long-Term Allograft Function. JASN 14, 1662–1668. https://doi.org/10.1097/01.ASN.0000066143.02832.5E

Harwood, R., Bridge, J., Ressel, L., Scarfe, L., Sharkey, J., Czanner, G., Kalra, P., Odudu, A., Kenny, S., Wilm, B., Murray, P., 2019. Murine models of renal ischaemia reperfusion injury: An opportunity for refinement using non-invasive monitoring methods (preprint). Physiology. https://doi.org/10.1101/2019.12.17.879742

Heinemann, F., Birk, G., Schoenberger, T., Stierstorfer, B., 2018. Deep neural network based histological scoring of lung fibrosis and inflammation in the mouse model system. PLoS One 13, e0202708. https://doi.org/10.1371/journal.pone.0202708

Hermsen, M., Bel, T. de, Boer, M. den, Steenbergen, E.J., Kers, J., Florquin, S., Roelofs, J.J.T.H., Stegall, M.D., Alexander, M.P., Smith, B.H., Smeets, B., Hilbrands, L.B., Laak, J.A.W.M. van der, 2019. Deep Learning–Based Histopathologic Assessment of Kidney Tissue. JASN 30, 1968–1979. https://doi.org/10.1681/ASN.2019020144

Hesketh, E.E., Czopek, A., Clay, M., Borthwick, G., Ferenbach, D., Kluth, D., Hughes, J., 2014. Renal Ischaemia Reperfusion Injury: A Mouse Model of Injury and Regeneration. JoVE 51816. https://doi.org/10.3791/51816

Hölscher, D.L., Bouteldja, N., Joodaki, M., Russo, M.L., Lan, Y.-C., Sadr, A.V., Cheng, M., Tesar, V., Stillfried, S.V., Klinkhammer, B.M., Barratt, J., Floege, J., Roberts, I.S.D., Coppo, R., Costa, I.G., Bülow, R.D., Boor, P., 2023. Next-Generation Morphometry for pathomics-data mining in histopathology. Nat Commun 14, 470. https://doi.org/10.1038/s41467-023-36173-0

Jayapandian, C.P., Chen, Y., Janowczyk, A.R., Palmer, M.B., Cassol, C.A., Sekulic, M., Hodgin, J.B., Zee, J., Hewitt, S.M., O'Toole, J., Toro, P., Sedor, J.R., Barisoni, L., Madabhushi, A., 2021. Development and evaluation of deep learning–based segmentation of histologic structures in the kidney cortex with multiple histologic stains. Kidney Int 99, 86–101. https://doi.org/10.1016/j.kint.2020.07.044

Kato, T., Relator, R., Ngouv, H., Hirohashi, Y., Takaki, O., Kakimoto, T., Okada, K., 2015. Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image. BMC Bioinformatics 16, 316. https://doi.org/10.1186/s12859-015-0739-1

Kshirsagar, A.V., Joshi, S.C., n.d. 2D IMAGE SEMANTIC SEGMENTATION FOR SELF-DRIVING CAR USING CONVOLUTION NEURAL NETWORK 10.

Lameire, N.H., Bagga, A., Cruz, D., De Maeseneer, J., Endre, Z., Kellum, J.A., Liu, K.D., Mehta, R.L., Pannu, N., Van Biesen, W., Vanholder, R., 2013. Acute kidney injury: an increasing global concern. The Lancet 382, 170–179. https://doi.org/10.1016/S0140-6736(13)60647-9

Lindsay, G.W., 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. Journal of Cognitive Neuroscience 33, 2017–2031. https://doi.org/10.1162/jocn_a_01544

Marechal, E., Jaugey, A., Tarris, G., Paindavoine, M., Seibel, J., Martin, L., Funes de la Vega, M., Crepin, T., Ducloux, D., Zanetta, G., Felix, S., Bonnot, P.H., Bardet, F., Cormier, L., Rebibou, J.-M., Legendre, M., 2022. Automatic Evaluation of Histological Prognostic Factors Using Two Consecutive Convolutional Neural Networks on Kidney Samples. Clinical Journal of the American Society of Nephrology 17, 260. https://doi.org/10.2215/CJN.07830621

Morawietz, G., Ruehl-Fehlert, C., Kittel, B., Bube, A., Keane, K., Halm, S., Heuser, A., Hellmann, J., 2004. Revised guides for organ sampling and trimming in rats and mice – Part 3. Experimental and Toxicologic Pathology 55, 433–449. https://doi.org/10.1078/0940-2993-00350

Niazi, M.K.K., Parwani, A.V., Gurcan, M., 2019. Digital Pathology and Artificial Intelligence. Lancet Oncol 20, e253–e261. https://doi.org/10.1016/S1470-2045(19)30154-8

Niazi, M.K.K., Tavolara, T.E., Arole, V., Hartman, D.J., Pantanowitz, L., Gurcan, M.N., 2018. Identifying tumor in pancreatic neuroendocrine neoplasms from Ki67 images using transfer learning. PLoS One 13, e0195621. https://doi.org/10.1371/journal.pone.0195621

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G.V., Krpalkova, L., Riordan, D., Walsh, J., 2020. Deep Learning vs. Traditional Computer Vision, in: Arai, K., Kapoor, S. (Eds.), Advances in Computer Vision, Advances in Intelligent Systems and Computing. Springer International Publishing, Cham, pp. 128–144. https://doi.org/10.1007/978-3-030-17795-9_10

Percy, D.H., Barthold, S.W., 2007. Pathology of laboratory rodents and rabbits, 3rd ed. ed. Blackwell Pub, Ames, Iowa.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation.

Scudamore, C.L., n.d. A Practical Guide to the Histology of the Mouse 250.

Sharkey, J., Ressel, L., Brillant, N., Scarfe, L., Wilm, B., Park, B.K., Murray, P., 2019. A Noninvasive Imaging Toolbox Indicates Limited Therapeutic Potential of Conditionally Activated Macrophages in a Mouse Model of Multiple Organ Dysfunction. Stem Cells International 2019, 1–13. https://doi.org/10.1155/2019/7386954

Sun, P., Liu, J., Li, W., Xu, X., Gu, X., Li, H., Han, H., Du, C., Wang, H., 2016. Human endometrial regenerative cells attenuate renal ischemia reperfusion injury in mice. J Transl Med 14, 28. https://doi.org/10.1186/s12967-016-0782-3

Tang, J., Yuan, F., Shen, X., Wang, Z., Rao, M., He, Y., Sun, Y., Li, X., Zhang, W., Li, Y., Gao, B., Qian, H., Bi, G., Song, S., Yang, J.J., Wu, H., 2019. Bridging Biological and Artificial Neural Networks with Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges. Advanced Materials 31, 1902761. https://doi.org/10.1002/adma.201902761

Wang, W., Faubel, S., Ljubanovic, D., Mitra, A., Falk, S.A., Kim, J., Tao, Y., Soloviev, A., Reznikov, L.L., Dinarello, C.A., Schrier, R.W., Edelstein, C.L., 2005. Endotoxemic acute renal failure is attenuated in caspase-1-deficient mice. American Journal of Physiology-Renal Physiology 288, F997–F1004. https://doi.org/10.1152/ajprenal.00130.2004

Xie, G., Chen, Tiange, Li, Y., Chen, Tingyu, Li, X., Liu, Z., 2020. Artificial Intelligence in Nephrology: How Can Artificial Intelligence Augment Nephrologists' Intelligence? Kidney Dis (Basel) 6, 1–6. https://doi.org/10.1159/000504600

Figures and Table

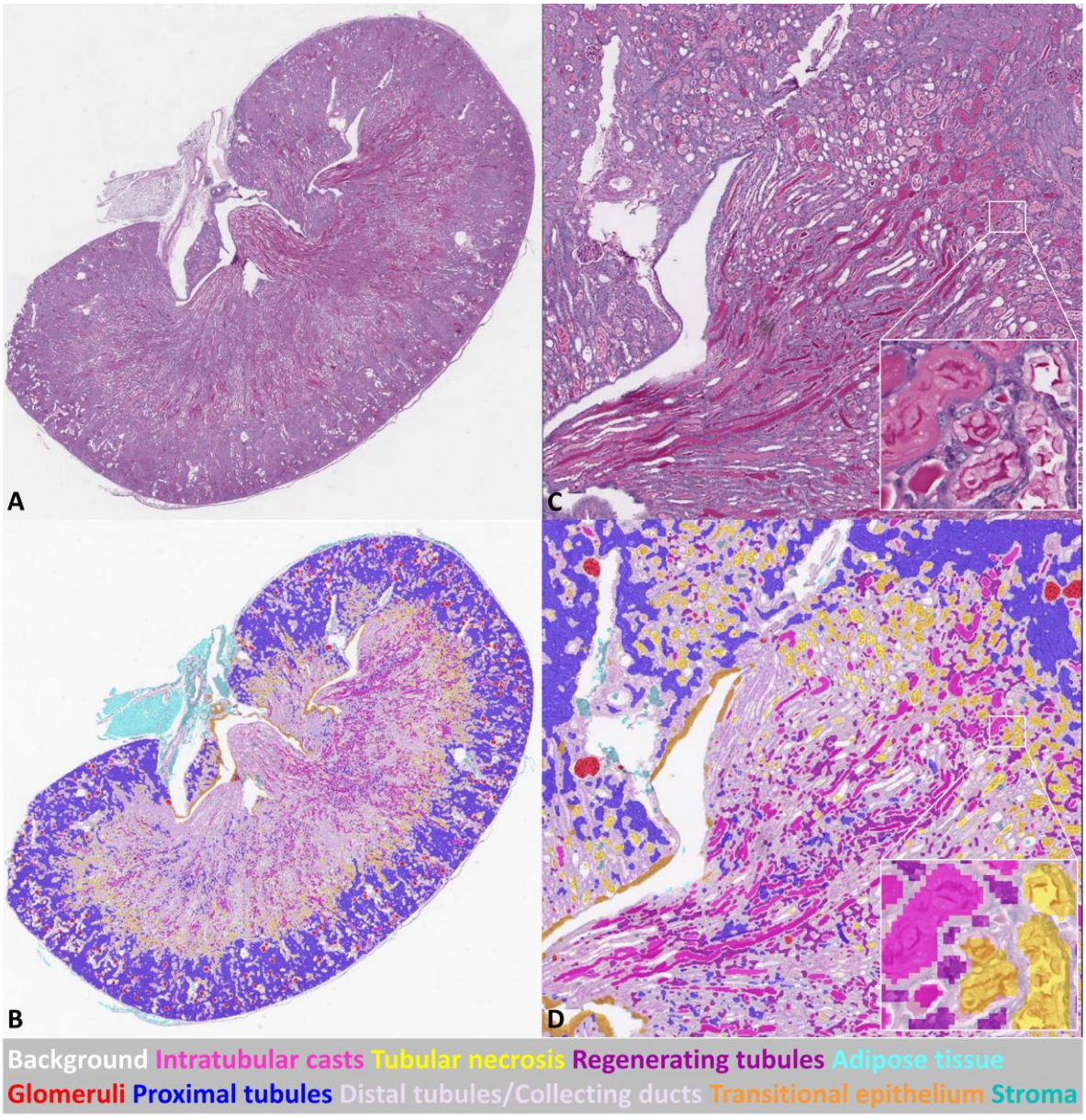

Fig. 1. CNN-based automated segmentation on WSI of an injured murine kidney on day 3 after IRI. (A) PAS stained WSI; (B) corresponding segmentation result of WSI in A; (C) High magnification PAS stained WSI, with further higher magnification area (inset); (D) corresponding segmentation result of WSI in C, with further higher magnification area (inset).

Fig. 2. Heat map confusion matrix for the CNN performance on the test set of mouse kidney IRI sections. The ground truth classes are given vertically (percentage of pixels), and the predicted classes (percentage of pixels) are shown on the horizontal axis. Examples of readings: 85% of all pixels labelled as "Tubular necrosis" were classified as "Tubular necrosis" by the DL model; 94% of all pixels labelled as "Glomeruli" were classified as "Glomeruli" by the DL model; 44% of the pixels labelled as "Regenerating epithelium" were misclassified with the "Distal tubules/Collecting ducts" class.

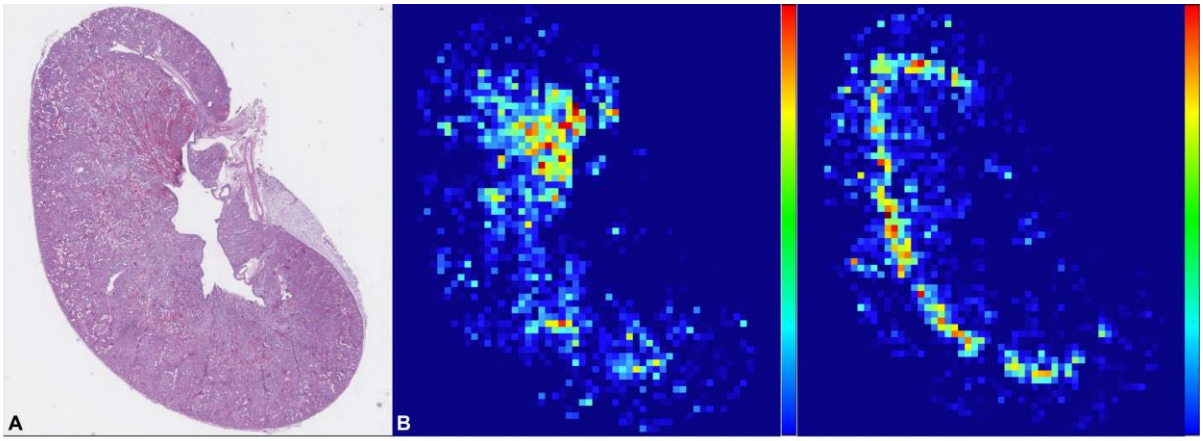

Fig. 3. PAS stained WSI of mouse kidney (A) and heatmaps of "Intratubular casts" (B) and "Tubular necrosis" (C). Percentage of area of each patch (512x512 pixels) occupied by pathological classes is represented as colours ranging from deep blue (0%) to red (100%) in order to spatially visualise the pathological classes within kidney parenchyma.

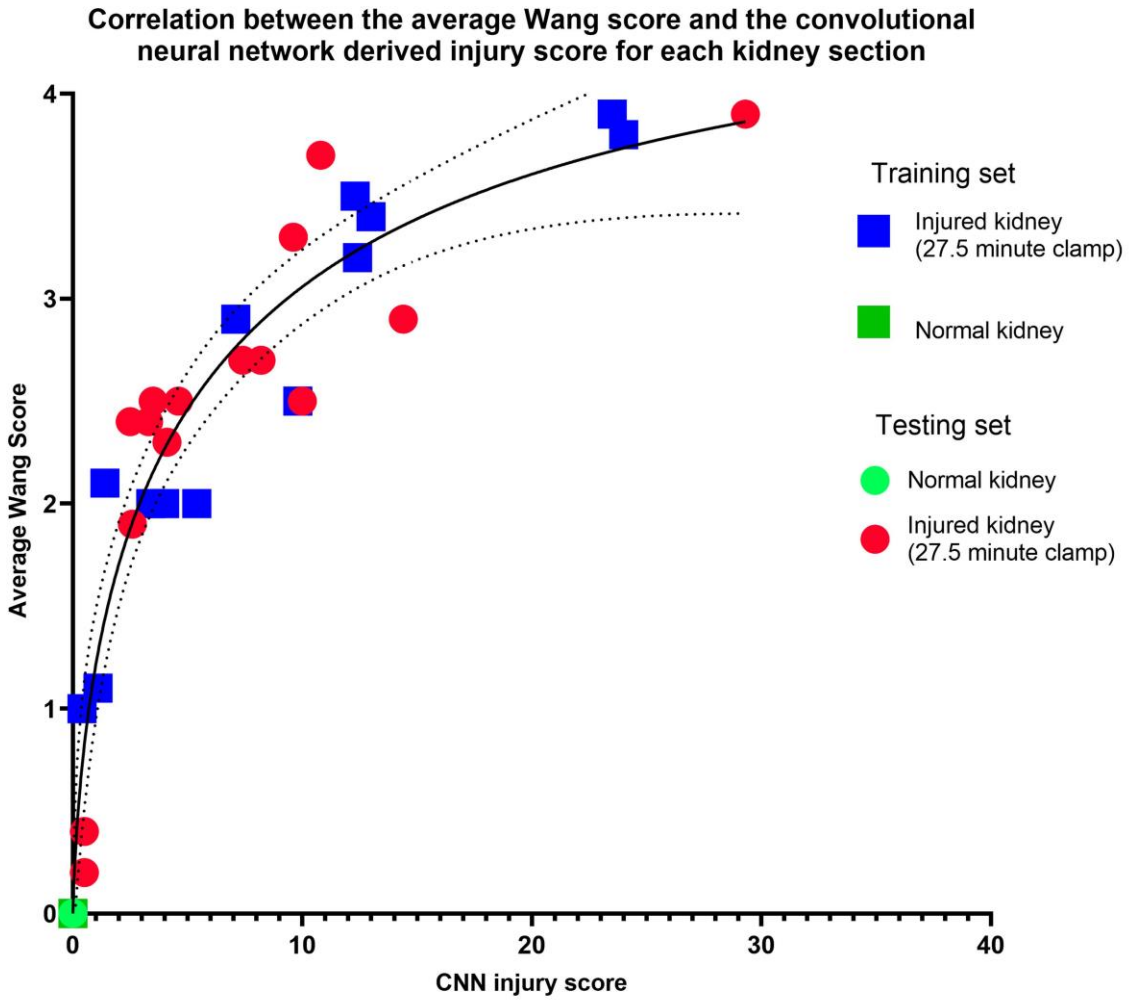

Fig. 4. Scatterplots visualising the correlation between the traditional scoring systems and CNN-based scoring per kidney section (N=34; Normal kidney results overlap at 0). The conventional method of scoring appears highly correlated to the CNN scoring, with the Spearman Correlation coefficient = 0.94 ($p<0.0001$).

Table 1. Accuracy parameters for CNN performance on the test set

Accuracy Parameter				
Class of interest	**Precision**	**True positive rate**	**Specificity**	**F1**
Background	1.00	0.97	1.00	0.98
Intratubular casts	0.87	0.70	1.00	0.78
Tubular necrosis	0.94	0.85	1.00	0.89
Regenerating epithelium	0.94	0.13	1.00	0.23
Adipose tissue	0.77	0.93	0.98	0.85
Glomeruli	0.99	0.94	1.00	0.97
Proximal tubules	0.97	0.88	1.00	0.92
Distal tubules/Collecting ducts	0.58	0.91	0.99	0.71
Transitional epithelium	0.85	0.94	1.00	0.89
Stroma	0.63	0.94	0.98	0.76

*Precision refers to the proportion of items classified as belonging to a particular class that are actually part of that class. True positive rate (Recall) represents the proportion of items correctly identified as belonging to a specific class out of all the items that truly belong to that class. Specificity is the metric that evaluates a model's ability to predict true negatives of each available category. F-score is a metric that combines precision and recall into a single measure. It is calculated by taking twice the product of precision and recall and dividing it by the sum of precision and recall.

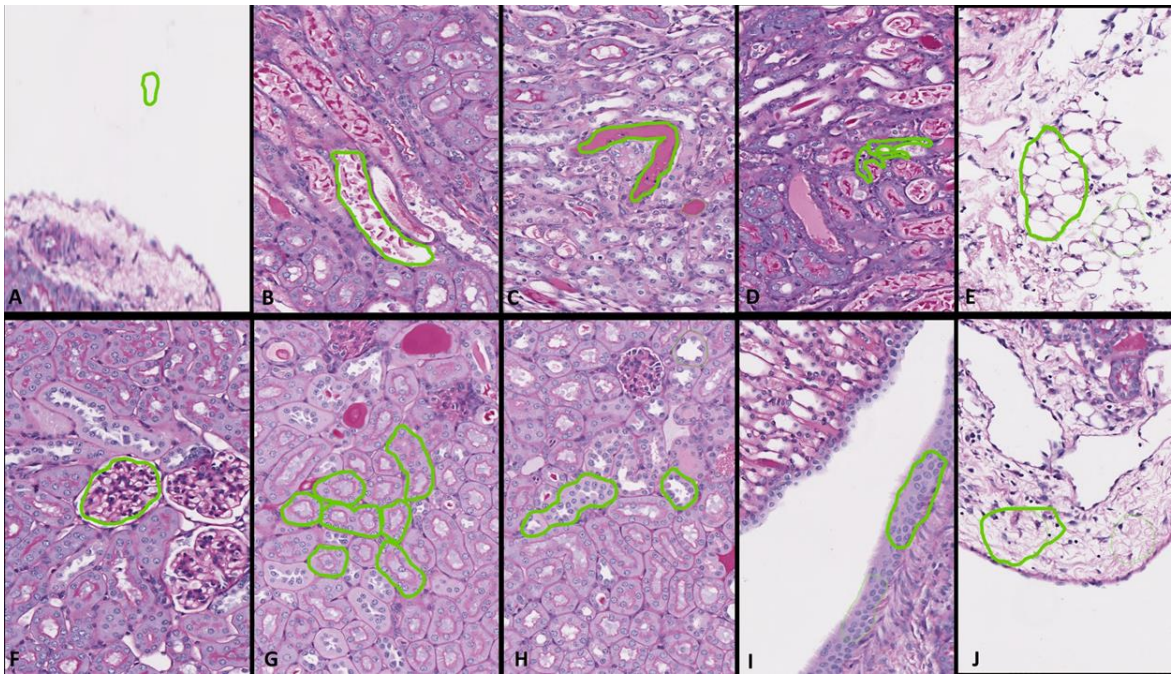

Fig. S1. Annotation procedure: A – Background, B – Tubular necrosis, C – Intratubular casts, D – Regenerating epithelium, E – Adipose tissue, F – Glomeruli, G – Proximal tubules, H – Distal tubules, I – Transitional epithelium, J – Stroma

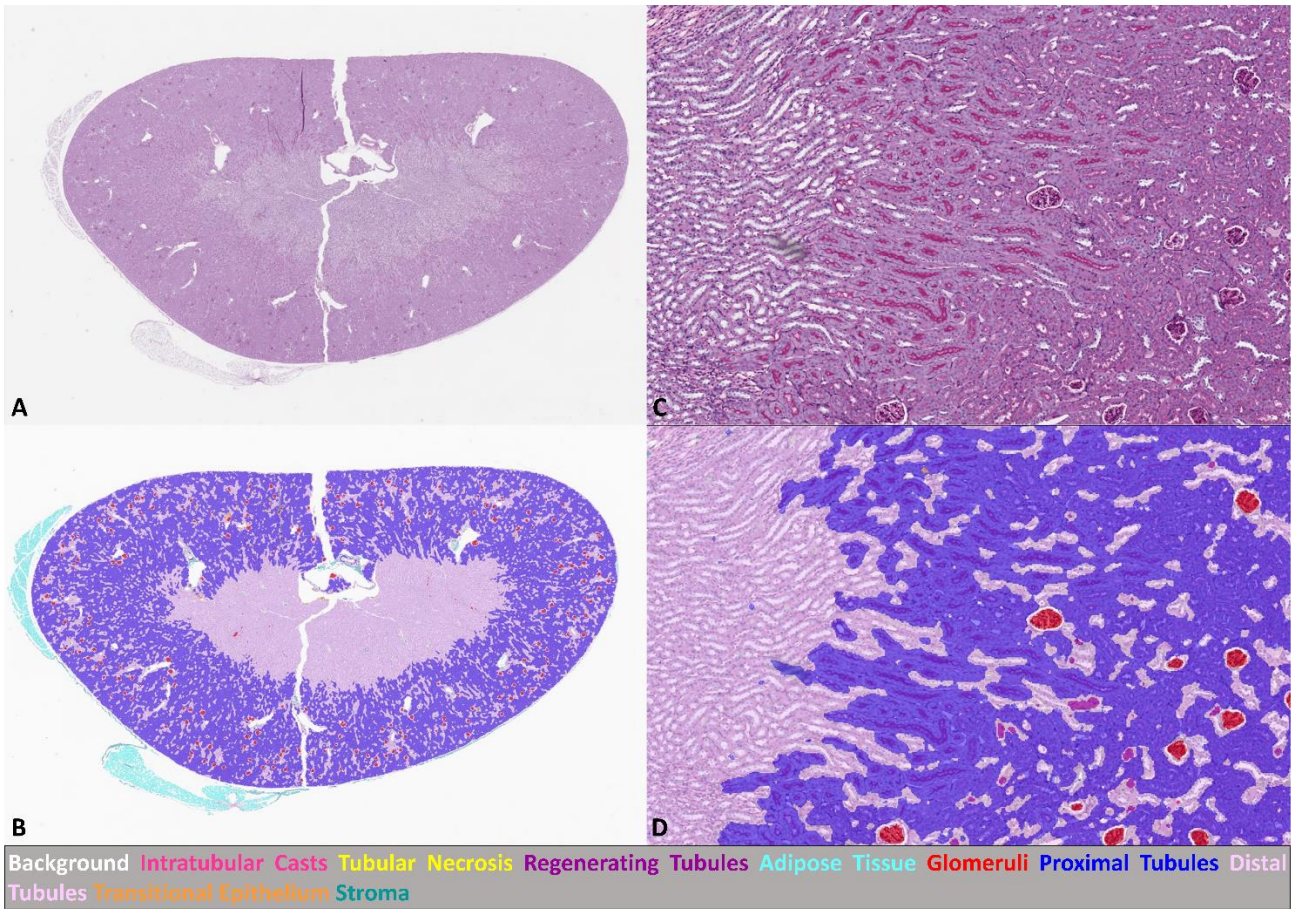

Fig. S2. Automated segmentation on WSI of a healthy murine kidney. (A) - full PAS stained WSI and its corresponding segmentation result (B); (C)- High magnification PAS stained WSI and its corresponding segmentation result (D).

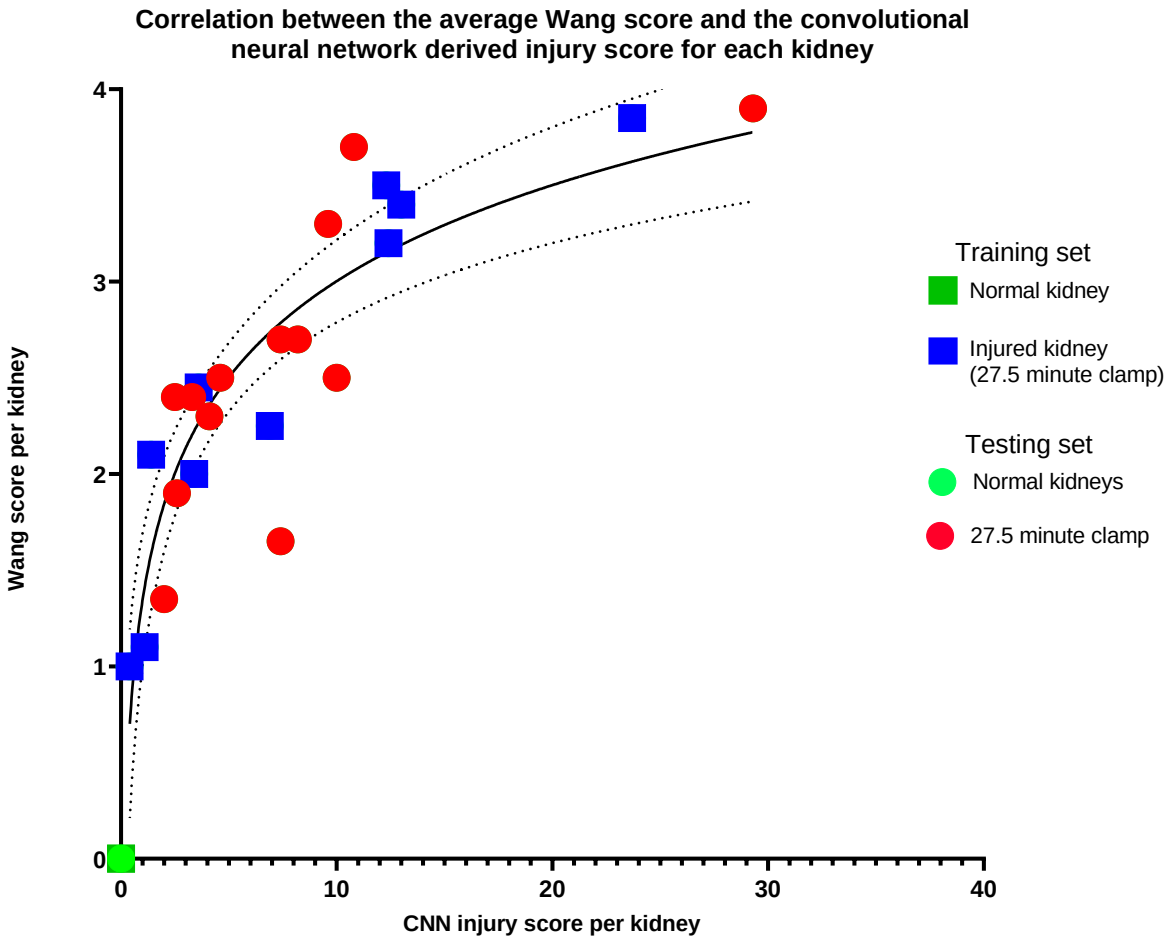

Fig. S3. Scatterplots visualising the correlation between the traditional scoring systems and CNN-based scoring per kidney. The conventional scoring method appears highly correlated to the CNN scoring, with the Spearman Correlation coefficient = 0.92 ($p<0.0001$).

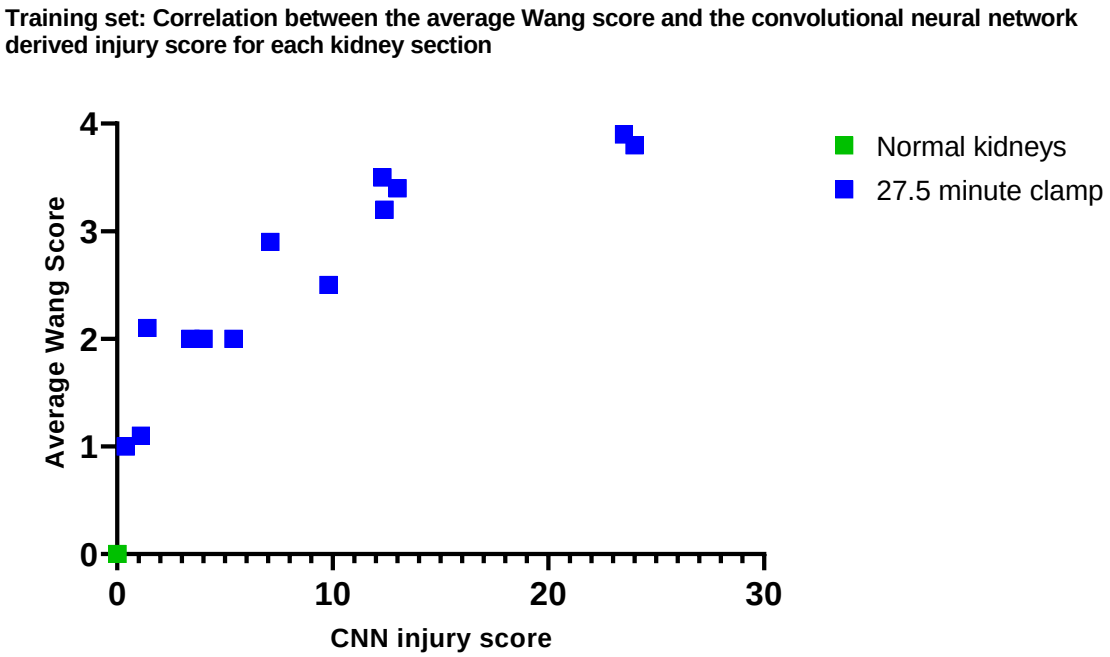

Fig. S4. Scatterplots visualising the correlation between the traditional scoring systems and CNN-based scoring per kidney section in the training set. The conventional scoring method appears highly correlated to the CNN scoring, with the Spearman Correlation coefficient = 0.95 ($p<0.0001$).

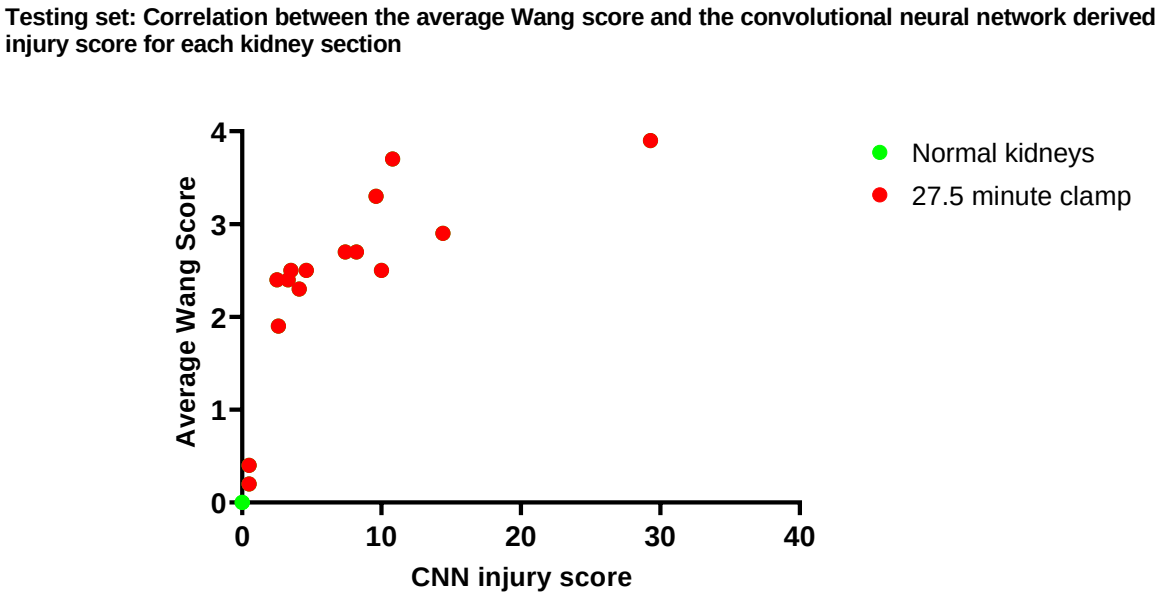

Fig. S5. Scatterplots visualising the correlation between the traditional scoring systems and CNN-based scoring per kidney section in the testing set. The conventional scoring method appears highly correlated to the CNN scoring, with the Spearman Correlation coefficient = 0.92 ($p<0.0001$).

Table S1. Quantitative information on ground truth data

Class	Number of annotations	Number of pixels
Background	288	76952097
Intratubular casts	288	2198198
Tubular necrosis	288	3033026
Regenerating epithelium	288	8459979
Adipose tissue	288	3248413
Glomeruli	288	3871080
Proximal tubules	288	3579512
Distal tubules/Collecting ducts	288	3858445
Transitional epithelium	288	8275538
Stroma	288	888184

Table S2. IRI damage scorings (test + train) performed by the DL model and the pathologist

Case	DL (CNN score)	Pathologist score (Wang modified)
1	1.10%	1.1
2	4.00%	2
3	9.80%	2.5
4	5.40%	2

5	7.10%	2.9
6	23.50%	3.9
7	24.00%	3.8
8	13.00%	3.4
9	12.40%	3.2
10	3.40%	2
11	0.40%	1
12	1.40%	2.1
13*	0.00%	0
14*	0.00%	0
15	12.30%	3.5
16	10.00%	2.5
17	0.50%	0.4
18	14.40%	2.9

19	0.50%	0.2
20	3.50%	2.5
21	2.50%	2.4
22	4.60%	2.5
23	8.20%	2.7
24	2.60%	1.9
25	7.40%	2.7
26	9.60%	3.3
27	3.30%	2.4
28	10.80%	3.7
29	29.30%	3.9
30	4.10%	2.3
31*	0.00%	0

Cases marked with an * - 2 sections per slide, scores represent the average of those 2 consecutive sections

Table S3. Overall model statistics

Mean Precision	0.85
Mean True Positive Rate	0.82
Mean Specificity	1
Mean F1	0.80

Table S4. Accuracy parameters for CNN performance on the training set

Accuracy Parameter				
Class of interest	**Precision**	**True positive rate**	**Specificity**	**F1**
Background	1.00	0.97	1.00	0.98
Intratubular casts	0.84	0.87	0.99	0.85
Tubular necrosis	0.99	0.84	1.00	0.91
Regenerating epithelium	0.71	0.46	1.00	0.56
Adipose tissue	0.90	0.91	0.97	0.91
Glomeruli	0.99	0.91	1.00	0.95
Proximal tubules	0.99	0.85	1.00	0.92
Distal tubules/Collecting ducts	0.77	0.91	0.96	0.83
Transitional epithelium	0.94	0.94	1.00	0.94
Stroma	0.72	0.96	0.95	0.83