Scientific Reports

https://doi.org/10.1038/s41598-026-61818-7

Article in Press

CYSTSCAN–PKD: a comprehensive pipeline for automatic cyst segmentation and counting on µCT scans from PKD animal models

Received: 11 May 2026

Accepted: 7 July 2026

Published online: 14 July 2026

Cite this article as: Mangili A., Arrigoni A., Sangalli F. *et al.* CYSTSCAN–PKD: a comprehensive pipeline for automatic cyst segmentation and counting on µCT scans from PKD animal models. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-61818-7

Andrea Mangili, Alberto Arrigoni, Fabio Sangalli, Stephanie Fest-Santini, Daniela Corna, Domenico Cerullo, Christodoulos Xinaris, Susanna Tomasoni, Maurizio Santini, Andrea Remuzzi & Anna Caroli

CYSTSCAN–PKD: a comprehensive pipeline for automatic cyst segmentation and counting on µCT scans from PKD animal models

Andrea Mangili[1], Alberto Arrigoni[1], Fabio Sangalli[1], Stephanie Fest-Santini[2], Daniela Corna[3], Domenico Cerullo[3], Christodoulos Xinaris[3], Susanna Tomasoni[3], Maurizio Santini[2], Andrea Remuzzi[4] and *Anna Caroli[1]

1. Bioengineering Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Bergamo, Italy.
2. Department of Engineering and Applied Sciences, University of Bergamo, Dalmine, Bergamo, Italy.
3. Department of Molecular Medicine, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Bergamo, Italy.
4. Department of Management, Information and Production Engineering, University of Bergamo, Dalmine, Bergamo, Italy.

Corresponding author:

Anna Caroli

Medical Imaging Laboratory, Department of Biomedical Engineering

Istituto di Ricerche Farmacologiche Mario Negri IRCCS

Via G.B. Camozzi 3, 24020, Ranica (BG)

anna.caroli@marionegri.it

Phone: +39 03545351

Fax: +39 0354535371

Abstract.

Autosomal dominant polycystic kidney disease (ADPKD) is a genetic disorder causing progressive renal cyst formation, increased kidney volume, and impaired function. The PCK rat is a preclinical model for investigating new treatments, requiring accurate quantification of total kidney volume (TKV), total cyst

volume (TCV), and cyst count. We propose an automated segmentation pipeline of kidneys and cysts on μCT scans of excised rat kidneys, followed by automated cyst counting.

For segmentation, a 3D U-Net ensemble was implemented using the nnU-Net framework with dual-channel input (raw μCT volumes, Sobel-filtered images). Models were trained on Dataset 1 subsets (D1, n = 5, 10, 15, 20), and evaluated on internal test set (n = 5) and independent external Dataset 2 (D2, n = 5). Segmentation achieved Dice Similarity Coefficients >0.99 for kidney and >0.98 for cysts on D1, with comparable performance on D2, exploring cross-scanner generalizability.

For cyst counting, a morphological algorithm based on 3D distance transform and peak detection was optimized via genetic algorithm on 10 samples and evaluated on independent 10-sample set. The automated method achieved low variability, with operator-algorithm agreement exceeding inter-operator agreement across all metrics, drastically reducing processing time.

The pipeline provides fast, accurate, and reproducible quantification of morphological biomarkers required for preclinical PKD research.

Keywords: Polycystic kidney disease (PKD); Cystogenesis; μCT imaging; Automatic Segmentation; U-Net; Cyst counting

1 Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is the most prevalent genetic kidney disorder [1–3]. This condition is characterized by the uncontrolled formation of multiple fluid-filled cysts within the renal tubules. The process of cystogenesis has serious consequences, including increased total kidney volume (TKV) and a corresponding reduction in parenchymal volume, which leads to a progressive decline in kidney function often culminating in end-stage renal disease (ESRD), requiring dialysis or transplantation [4,5].

Despite advancements in understanding the pathophysiology of ADPKD, effective therapies capable of preventing cyst formation or halting disease progression remain elusive [6]. This gap is linked to three critical needs: i) novel effective drugs, ii) reliable animal models for investigating new pharmacological strategies, and iii) robust biomarkers for monitoring disease progression and evaluating therapeutic responses. Addressing these needs is essential for supporting preclinical and clinical research on ADPKD, and ultimately improving long-term outcomes in ADPKD patients.

Animal models are essential for elucidating the complex pathophysiological mechanisms that drive cystogenesis in ADPKD and for evaluating candidate therapies aimed at slowing disease progression. In particular, the PCK rat is a genetically inherited model of polycystic kidney disease (PKD) that develops spontaneously in Sprague-Dawley (SD) rats. Although it follows an autosomal recessive inheritance pattern, it exhibits many features resembling human ADPKD [7].

Among available biomarkers, TKV is widely recognized as a significant indicator of disease progression and treatment efficacy [8,9]. Recent studies also suggest that total cyst volume (TCV) may be more informative, given the strong correlation between cyst burden and renal functional decline [10,11].

Three-dimensional micro-computed tomography (μCT) imaging has emerged as a powerful tool for accurately quantifying TKV and TCV in pre-clinical models. However, manual segmentation of renal and cystic structures from μCT datasets remains extremely time-consuming and operator-dependent, limiting its scalability [12,13].

Beyond volumetric biomarkers, individual cyst enumeration provides complementary insights into disease mechanisms [14]. Cyst count may also be more sensitive than volume for monitoring disease progression and detecting early treatment effects, particularly for therapies targeting cystogenesis rather than growth inhibition. Manual cyst counting in polycystic kidneys poses significant challenges because cysts tend to be densely clustered, may have irregular protruding shapes, and often touch or branch into complex three-dimensional patterns [15–17]. Manual counting is time-consuming and prone to high inter-operator variability, particularly when distinguishing individual cysts in clustered regions. These limitations have hindered the adoption of cyst counting as a routine quantitative endpoint in preclinical studies. Despite the potential value of automated counting, to our knowledge, no such methods exist in μCT-based preclinical research.

In recent years, deep learning (DL), and particularly convolutional neural networks (CNNs), have substantially advanced medical image segmentation tasks. Notably, the U-Net architecture and its 3D variants have achieved high performance across diverse applications, from organ to tumour segmentation [18–21]. When considering polycystic kidney image analysis, several studies have applied DL to medical image data: Sharma et al. [13] achieved high accuracy for kidney segmentation on CT scans, while Kline et al. [11] pioneered instance segmentation for individual cyst identification on MRI. However, in the context of μCT imaging and preclinical research, automated segmentation of polycystic rat kidneys has been addressed in only one published study to date, by Rombolotti et al. [12], who adopted a 2D U-Net approach, potentially losing volumetric context. Importantly, μCT introduces additional challenges: the

higher spatial resolution increases structural complexity and the number of small cystic features to resolve, and preclinical studies often require detecting subtle, treatment-induced morphological differences, which demands particularly robust and sensitive segmentation.

To address these challenges, we developed and validated an automated analysis pipeline for μCT scans from polycystic rat kidneys that combines 3D U-Net–based semantic segmentation of kidneys and cysts with genetic algorithm–optimized morphological cyst counting.

Specifically, this study aims to: (1) develop an automatic, accurate, and robust 3D U-Net ensemble model for kidney and cyst volume segmentation, comparing different training sample sizes and evaluating cross-scanner generalizability; and (2) develop an automated cyst counting method operating on cyst segmentations, validated against independent manual counts from two operators to quantify inter- and intra-operator variability.

The proposed integrated approach constitutes the first comprehensive automated image analysis pipeline designed to deliver all morphological imaging biomarkers required for preclinical PKD studies.

2 Methods

2.1 Animal populations and sample preparation

Forty-one male PCK rats and two male Sprague-Dawley (SD) rats of 3 weeks of age were obtained from Charles River Laboratories and housed in a specific pathogen-free facility at constant temperature, under a 12:12 h light/dark cycle, with free access to food and water. At 4 weeks of age, eight PCK and one SD rats were anesthetized with isoflurane and maintained under stable anesthesia. The left kidney of each rat was then perfused with a radio-opaque silicone polymer (Microfil MV-122; Flow Tech, Carver, MA, USA) according to a previously described protocol [22]. Excised kidneys were fixed and stored in 10% formalin in a centrifuge tube. At 14 weeks of age, the remaining thirty-three PCK and one SD rats underwent the same procedure.

All procedures involving animals were performed in accordance with institutional guidelines in compliance with national (D.L.n.26, March 4, 2014), and international laws and policies (directive 2010/63/EU on the protection of animals used for scientific purposes). This study was approved by the Institutional Animal Care and Use Committees of Istituto di Ricerche Farmacologiche Mario Negri IRCCS and by the Italian Ministry of Health (approval number 991/2023-PR).

2.2 Software

The image processing pipeline proposed in this study was developed using MATLAB (vR2023b, https://www.mathworks.com/) and Python (v3.8.18; https://www.python.org/), exploiting pertinent libraries and additional software, including 3D Slicer (v5.6.2; www.slicer.org) and ImageJ/Fiji (v.1.54p; http://imagej.org).

2.3 µCT image datasets

µCT images were used for the pipeline development and internal evaluation (Dataset 1) and for external testing, to assess the generalizability of the segmentation model on images acquired with a different scanner (Dataset 2), respectively. Dataset-specific image acquisition, pre-processing and ground truth details are reported hereafter.

2.3.1 Dataset 1 (D1)

Dataset 1 comprised µCT scans from thirty-eight rats, sacrificed at 4 weeks (8 PCK and 1 SD control) or 14 weeks (28 PCK and 1 SD). Of these, twenty-five (n = 25) were allocated to the kidney and cyst segmentation task and twenty (n = 20) to the cyst counting task, with seven PCK samples overlapping between the two subsets.

Image acquisition. µCT images were acquired at the University of Bergamo with a 9×9×9 µm^3 isotropic resolution using an open-type X-ray source operating at 120 kV and 20 μA_{target} in high-power mode. The detector was a 16-bit CMOS with an acquisition time of 1.2 s, and 3200 projections were acquired per tomography. During acquisition, the projections were corrected for detector charge accumulation, the so-called *"dark-field"*, and were normalized by *"bright-field"* correction. The latter characterizes the non-uniform response of the radiography acquisition system due to variations in detector sensitivity and X-ray source flux density. The normalized projections are then fed into the Feldkamp-Davis-Kress (FDK) reconstruction algorithm in VGSTUDIO MAX to obtain the three-dimensional dataset. The specified isotropic resolution was determined with a metrological calibration. Calibration procedures are detailed in Santini et al. [23]. Excised kidneys were scanned in the 10% formalin centrifuge tube to minimize beam-hardening effects

Image pre-processing. To reduce the burden of manual segmentation and ensure feasibility in terms of memory management and computational capacity, all scans were resampled to an isotropic resolution of

27×27×27 μm^3. The preprocessed scans had median matrix dimensions of 742×777×970 voxels. Scans were stored as volumetric images in compressed NIFTI format.

Training/test split - kidney and cysts segmentation. All 38 D1 scans were retained for analysis; no samples were excluded due to image quality, scanner, or acquisition issues, or lack of annotation. The segmentation dataset comprised the first 25 consecutive eligible D1 scans. Five of these were reserved as a fixed hold-out test set, selected to span the range of disease burden: (i) two 14-week-old cases with the highest and lowest total cyst volume (TCV); (ii) two 14-week-old cases with the largest and smallest total kidney volume (TKV); and (iii) one 4-week-old case with the highest TCV. The remaining 20 samples formed the training pool, from which four nested training subsets of 5, 10, 15, and 20 samples were constructed to investigate the impact of training-set size (see **Table 1**).

SD control samples were distributed across training sets as follows: one SD sample was included in each of the 5- and 10-sample training subsets, and both SD samples were included in the 15- and 20-sample subsets. The ratio of 4-week to 14-week PCK samples was maintained at approximately 1:4 across all training subsets and the test set, consistent with their distribution in the source population.

Each training set was purposely constructed to capture the variability observed in μCT images, arising from differences in age and pathological status, and to account for representative challenging cases, including i) samples with poorly perfused Microfil regions, which are challenging even in the case of manual segmentation; ii) samples presenting air bubbles; iii) samples with presence of Microfil within cysts and/or the renal papilla, which are difficult to distinguish from vasculature due to a similar intensity but spherical shape; and iv) at least one control SD sample.

Training/test split – cyst counting. Twenty μCT samples from Dataset 1 were used to develop and validate the cyst-counting algorithm. This set comprised D1 scans not assigned to the segmentation dataset, together with seven samples in the segmentation training set; none of these seven overlapping samples belonged to the fixed segmentation hold-out test set. The dataset was split into:

- Optimization dataset: 10 samples used for parameters optimization.
- Evaluation dataset: 10 samples used for assessing the cyst-counting algorithm performance and the inter-operator variability (see **Table 1**).

All twenty samples were PCK rats, as these animal models develop multiple cysts and thus provide a suitable segmentation mask for algorithm development and evaluation. The optimization dataset comprised exclusively 14-week-old animals, whereas the ten evaluation samples included five 4-week-old and five

14-week-old PCK rats. The per-sample allocation of all scans across the segmentation and counting analyses is reported in **Supplementary Table S1** online.

Ground truth - kidney and cysts segmentation. Initially, two experienced operators (Operator 1, AM; and Operator 2, FS) manually segmented kidneys and cysts using the Segment Editor module available in 3D Slicer. Manual tracing was performed on a total of 400 slices from a PCK sample and 200 slices from a control SD sample. In the vascular pole region, where the kidney contour is irregular, the boundary was completed by drawing a polygon connecting the visible extremes of the renal lobes.

We then trained two separate 2D U-Net models using the preliminary manual segmentations: one for kidney segmentation and one for cyst segmentation. Input images and corresponding masks underwent on-the-fly data augmentation to increase dataset size and variability, including random rotations (up to 360°), random-sized cropping, and horizontal flipping, while preserving label integrity through nearest-neighbour interpolation. The kidney model was trained for 50 epochs, whereas the cyst model was trained for 100 epochs. For both networks, a Shuffle-Split validation scheme was adopted, with an 80/20 train/validation split randomly resampled at each epoch. Training was performed with the Adam optimizer (learning rate = 1e-4), a batch size of 6, and a sigmoid activation in the output layer. Training the models required approximately one hour each.

The trained 2D U-Nets were then used to generate kidney and cyst segmentations for all 25 samples in Dataset 1 (D1), which were reviewed and manually refined by the two operators in an alternating sequence: for each sample, one operator edited the preliminary mask and the other revised it, yielding the final ground-truth annotations. Manual refinement was required on virtually all slices. We also quantified the correction burden by comparing the preliminary 2D U-Net masks against the final refined masks on 10 cases (the 5 D1 test samples, and the 5-sample training subset): kidney DSC was 0.9858 ± 0.0085 and cyst DSC was 0.7629 ± 0.2905, confirming that manual revision was extensive across cases.

Ground truth – cyst counting. Manual cyst counting was performed by Operator 1 alone on the optimization dataset and, independently, by both Operator 1 and Operator 2 on the evaluation dataset. Cyst counts were performed using ground-truth segmentation masks. Operators used 3D Slicer and were blinded to each other's counts and to the automated results. For the evaluation dataset, ground truth (GT) was defined as the average of the two operator counts, since inter-operator discrepancies typically occurred across samples rather than following a systematic pattern.

2.3.2 Dataset 2 (D2)

Dataset 2 included five µCT scans from PCK rats sacrificed at 14 weeks of age.

Image acquisition. µCT images were acquired at Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Bergamo, Italy, using a µCT SkyScan 1076 (Bruker, Belgium) at an isotropic resolution of 9×9×9 µm3. Acquisition parameters were as follows: source voltage of 55 kV, source current of 180 µA, 0.5 mm Al filter, exposure time of 1000 ms, and a rotation step of 0.5° (395 images). Volumetric images were reconstructed using a back-projection algorithm implemented in the NRecon software (Bruker, Belgium). Acquisition parameters for both datasets are compared in **Supplementary Table S2** online.

Image pre-processing. As in D1, the 3D images were resampled to an isotropic resolution of 27×27×27 μm^3. The preprocessed scans had median matrix dimensions of 653×653×938 voxels. A denoising step was also applied to D2 images using bilateral filtering in ImageJ/Fiji [24] to mitigate noise in this dataset. Images were stored in compressed NIFTI format.

Ground truth - kidney and cysts segmentation. Kidney and cyst ground-truth segmentations for these scans were generated semi-automatically by Operator 2 following the same procedure used for D1. The resulting masks were subsequently reviewed by Operator 1.

2.4 Kidney and Cysts Segmentation Ensemble Models

2.4.1 Development

Four 3D U-Net models were trained for semantic segmentation of kidneys and cysts using training subsets of 5, 10, 15, and 20 samples from D1, in order to assess the impact of training set size on segmentation accuracy. Training was performed with the nnU-Net automated deep learning framework (https://github.com/MIC-DKFZ/nnUNet), which automatically adapts the network architecture, preprocessing steps, and training hyperparameters to the specific characteristics of the dataset [25].

The model input consisted of i) the 3D µCT volumes, with a resolution of 27x27x27 μm^3; ii) the same 3D volumes filtered with the Sobel operator to enhance edges information; and iii) the pertinent ground truth segmentations of kidneys and cysts used as supervision target. The model's output included the predicted mask of kidneys and cysts and a voxel-wise probability map. To assess the contribution of the Sobel channel, we compared models trained on the 5-sample subset with the single-channel input (raw µCT only) versus the dual-channel input (raw µCT + Sobel).

Following the nnU-Net protocol, the 5-fold cross-validation scheme was adopted to (i) train a fold-wise ensemble for each training subset and (ii) obtain a robust estimate of segmentation performance. In each fold, 20% of the available training cases were held out for validation. Models were trained for 1000 epochs with an initial learning rate of $1e^{-3}$ and weight decay of $1e^{-5}$. The network architecture and training configuration relied on the nnU-Net default settings, consistent with evidence that careful implementation and optimization of the training pipeline are often more critical for state-of-the-art performance than ad hoc architectural modifications [25,26]. Each epoch used a batch size of 2 and a patch size of 160x112x128 voxels, with an approximate training time of 80 seconds per epoch and about 20 hours to train an entire fold. Following nnU-Net defaults, training used a combined Dice and cross-entropy loss.

Inference on the test sets was performed using an ensemble comprising the 5-fold-specific models. For each voxel, the final label (kidney, cyst, or background) was assigned by probability map averaging across the individual models. Inference on a new sample required approximately 30 minutes per case. **Figure 1** summarizes the 3D U-Net training pipeline and the ensemble-based inference workflow.

2.4.2 Evaluation

Cysts and kidney volumes were quantified on test sets-related ground truths and model predictions using the Segment Statistics tool in 3D Slicer. The performance of the ensemble models was evaluated by comparing the predicted kidney and cyst labels with the ground truth masks, using the Dice-Sørensen coefficient (DSC), Intersection over Union (IoU) scores, and the 99th percentile of the Hausdorff Distance (HD99) was used over the more commonly used HD95, given the large number of voxels belonging to each label. Bland-Altman and linear regression analyses were used to assess the agreement between volumes derived from predicted segmentations and ground-truth masks for both kidneys and cysts.

The benefit of incorporating the Sobel auxiliary channel was assessed by comparing the segmentation performance of two versions of the ensemble model, both trained on the 5-sample subset, with and without the additional input channel, on each sample of the D1 test set. Average Surface Distance (ASD) and Normalized Surface Distance (NSD) were additionally computed for both models. ASD is the mean bidirectional nearest-neighbour distance between predicted and ground-truth surface voxels (mm); NSD is the proportion of surface points (from both prediction and reference) within a tolerance τ of the opposite surface (τ = 0.1 mm, approximately 3.4 voxels at the 29.2 μm analysis resolution). Surface voxels were identified by morphological erosion, and distances were computed via Euclidean distance transform.

The impact of the training set size was assessed by comparing the segmentation performance of the four ensemble models (dual-channel) trained with different training sample sizes (n = 5, 10, 15, and 20) on the

D1 test set. In addition to the model-derived evaluation metrics, the statistical differences among the four ensembles were assessed, for each label separately, using the Friedman test.

The generalizability of the four ensemble models (dual channel) trained with different training sample sizes was finally evaluated on the D2 test set, consisting of images acquired with a scanner different from the one used to acquire training images.

All statistical analyses were conducted using the R software (R Core Team, Vienna, Austria, v4.3.3).

2.5 Automated Cyst Counting Algorithm

2.5.1 Development

We developed an automated cyst-counting algorithm operating from binary cyst masks. First, a 3D Euclidean distance transform (DT) was computed within the mask and smoothed to reduce spurious peaks. Candidate cyst centers were then identified as local maxima in the smoothed DT. Finally, morphology-based rules were applied to handle cysts separation and counting.

The cyst-counting pipeline is governed by six morphological parameters controlling peak detection on the DT map and the subsequent post-processing rules. Briefly, σ sets the Gaussian smoothing applied to the DT to suppress spurious local maxima, while T defines the adaptive threshold used to retain peaks based on local distance statistics. Cysts are classified as small or large using a volume threshold V_{thr} (in voxels), after which size-specific distance constraints are applied: d_S and d_L define the minimum inter-peak distance required to count two peaks as distinct cysts in small and large cysts, respectively, whereas d_M specifies the maximum distance for merging closely spaced peaks within large cyst regions to mitigate overcounting in clustered areas. **Figure 2** shows the algorithm workflow.

Algorithm 1: CYST COUNTING ALGORITHM

Parameters and GA optimization bounds

Symbol	Description	Bounds
σ	Standard deviation of the Gaussian kernel applied to the distance transform.	[0.01, 1.5] voxels [0.00027, 0.0405] mm
T	Adaptive threshold multiplier for peak selection	[0.1, 3.0]
V_{thr}	Volume threshold (voxels) separating small cysts from large cysts	[10.000, 150.000] voxels [0.197, 2.952] mm^3
d_S	Minimum peak separation for small cysts	[1, 15] voxels [0.027, 0.405] mm
d_M	Maximum distance for merging peaks within large cysts	[1, 15] voxels [0.027, 0.405] mm
d_L	Minimum peak separation for large cysts	[10, 100] voxels [0.270, 2.700] mm

Procedure

Input: Binary cyst mask M

Step 1. Extract cyst voxels from M. If empty, return 0.

Step 2. Compute the 3D Euclidean distance transform DT of the cyst region.
Normalize DT to [0, 1].

Step 3. Apply per-slice adaptive histogram equalization (CLAHE) to DT.
Rescaling the output back to each slice's original intensity range.

Step 4. Apply 3D Gaussian smoothing (σ) to produce smoothed map S.

Step 5. Identify local maxima in S using 26-connectivity (3x3x3 neighborhood).

Step 6. Extract connected components from M using 26-connectivity to identify individual cyst regions.

Step 7. Initialize total count $N = 0$.

Step 8. For each connected component C:

a. Classify size: if $|C| > V_{thr}$, set $d_{min} = d_L$ and enable merging;
Otherwise set $d_{min} = d_S$ and skip merging.

b. Compute adaptive thresold:
$\boldsymbol{\tau} = mean(S[C]) + T \cdot std(S[C])$

c. Select valid peaks:
Voxels in C that are local maxima in S and satisfy $S \geq \boldsymbol{\tau}$.
If none, continue.

d. Greedy suppression:
Iterate over valid peaks in index order;
For each surviving peak p, suppress all peaks within Euclidean distance d_{min}.
Lower index is retained on ties.

e. (Large cysts only)
Merging: if $||p - q|| < d_M$, remove q.

f. Add surviving peakl count to N.

Output: Return estimated cyst count *N*.

2.5.2 Optimization

Building on this workflow, cyst-counting algorithm parameters were optimized using a genetic algorithm (GA) approach on the 10-sample optimization dataset derived from D1. The six parameters, initially set empirically, were tuned using the GA to minimize counting errors.

The fitness function for the optimization (1) was designed to minimize the Mean Absolute Percentage Error (MAPE) between predicted and ground truth cyst counts. MAPE was chosen as the primary optimization metric because it provides scale-independent assessment of prediction accuracy, preventing high-burden samples from dominating the optimization process. The fitness function is defined as:

$$Fitness = MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{PC_i - GT_i}{GT_i}\right| \times 100 \qquad (Eq.\ 1)$$

where n represents the number of training samples, PC_i is the automated predicted count, and GT_i is the ground truth count for sample i.

The GA used a population size of 50 with tournament selection (tournament size = 3), adaptive feasible mutation (mutation rate = 0.95), elitism (2 individuals), a crossover fraction of 0.5, and a migration fraction of 0.15. Optimization ran for up to 100 generations, with early stopping after 10 consecutive stall generations. To prevent premature convergence, a diversity-injection strategy was implemented: if stagnation persisted for 5 consecutive generations, 50% of the non-elite population was reinitialized with random parameter values within the predefined bounds. To account for the stochastic nature of the GA, optimization was repeated across five independent runs.

In addition to MAPE, we reported complementary metrics informing on the optimization process: Mean Percentage Error (MPE) to assess directional bias; Mean Absolute Error (MAE) and its standard deviation to quantify absolute counting accuracy and variability; Root Mean Square Error (RMSE) to penalize large deviations; and the coefficient of determination (R^2) to evaluate goodness-of-fit between predicted and reference cyst counts.

2.5.3 Evaluation

After optimization, the automated algorithm (hereafter called Auto) was evaluated on the 10-sample independent evaluation dataset from D1 using:

- Bland-Altman and linear regression analyses, to assess the agreement and association i) between individual Operators and Auto cysts counting, ii) across operators, iii) between manual or Auto cyst counting and the consensus ground truth reference, defined as the average of the two operator counts.
- Performance metrics: coefficient of variation (CV), MAE, RMSE, Concordance Correlation Coefficient (CCC) with 95% CI via bootstrap (1000 iterations).
- Intraclass Correlation Coefficient, using a two-way random effects model with absolute agreement, ICC (2,1), for inter-operator and operator-algorithm agreement.

All evaluations were conducted using the R software (R Core Team, Vienna, Austria, v4.3.3).

2.5.4 Sensitivity Analysis

To assess the robustness of the optimized morphological parameters, we performed sensitivity analysis combining global Morris screening with a local one-at-a-time (OAT) perturbation analysis. Both Morris screening and OAT analyses were performed on the 10-sample optimization dataset.

Morris Screening. We applied Morris Elementary Effects method [27] to identify influential parameters across the entire parameter space. Following the improved sampling strategy proposed by Campolongo et al. [28], we generated $M = 1000$ candidate trajectories in the normalized unit hypercube using the following sampling step Δ (2):

$$\Delta = \frac{p}{2(p-1)} \qquad (Eq.\ 2)$$

with $p = 6$, and $r = 20$ trajectories selected with maximum spread based on pairwise geometric distances. Each trajectory was linearly transformed into the actual parameter ranges (±30% around optimal values), resulting in 140 model evaluations. For each parameter i, we computed: μ_i^* (mean absolute elementary effect), quantifying overall importance; μ_i (mean elementary effect), indicating sign consistency ($|\mu| \approx \mu^*$ suggests monotonic behavior, while $|\mu| << \mu^*$ indicates non-monotonic effects); and σ_i (standard deviation), revealing interactions and non-linearities [28].

Local Analysis. We performed a detailed OAT analysis on all six parameters, varying each one from -30% to +30% around its optimal value in 21 equally spaced steps while keeping the others fixed (126 evaluations). We also evaluated a 9×9 grid for the two most influential parameters to visualize their joint effect (81 evaluations). This combination of global screening and local analysis addresses the well-documented limitations of using OAT in isolation [29], while maintaining computational feasibility for this expensive model (~90 seconds per evaluation, ~7.5 hours total). All evaluations were performed using MATLAB with 8 parallel computing workers.

3 Results

Results are organized into two parts: (i) kidney and cyst segmentation by the ensemble model, including an analysis of the effect of an auxiliary edge-enhancing channel, the impact of training-set size, and external cross-scanner testing; (ii) automated cyst counting, including the evaluation of the agreement with dual-operator manual counts, and sensitivity analysis of the cyst-counting parameters.

3.1 Kidney and Cyst Segmentation Ensemble Models

The ensemble models worked successfully on all validation and test cases (representative examples can be found as **Supplementary Fig. S1** online). Inference took approximately 30 minutes on an Ubuntu 22.04 system with 256GB of RAM, an Intel Xeon 24-core processor running at 2.50 GHz, and an NVIDIA RTX A4500 GPU with 20GB of memory.

Impact of the Sobel channel. Supplementary Table S3 online reports the DSC, IoU, ASD and NSD segmentation performance scores achieved by the 3D ensemble models trained on the 5-sample D1 subset, with or without adding Sobel-filtered images as second input channel, for each sample of the D1 test set. Examples of masks obtained from the dual-channel model and the one-channel model, along with the ground-truth segmentation, can be found as **Supplementary Fig. S2** online.

For the dual-channel model, Bland–Altman analysis yielded a mean relative difference (bias) of 0.31% ± 0.56% for TKV, and −0.44% ± 0.42% for TCV. For the single-channel model, the corresponding bias was 0.24% ± 0.58% for TKV and −0.26% ± 0.74% for TCV (see **Supplementary Fig. S3** online). Linear regression showed near-perfect correlation for both models, with $R^2 > 0.999$ for TKV and TCV.

Impact of the training set size. Table 2 summarizes the DSC, IoU, and the HD99 segmentation performance metrics of the four ensemble models trained with different training sample sizes, for the kidney and cyst labels separately.

Segmentation metrics demonstrated high performance across all models. The Friedman test found no significant differences in kidney DSC, IoU, or HD99 (p = 0.12, 0.12, and 0.071, respectively). For cyst segmentation, significant differences were found in DSC and IoU (χ^2 = 12.60, p = 0.006) and HD99 (p = 0.011).

Bland–Altman analysis (see **Figure 3**) yielded the following mean relative differences (bias ± SD) between automated and ground truth segmentations:

- Training sample size n = 5: TKVs: 0.31% ± 0.56%, TCVs: –0.44% ± 0.42%;
- Training sample size n = 10: TKVs: 0.31% ± 0.67%, TCVs: –0.37% ± 0.61%;
- Training sample size n = 15: TKVs: 0.32% ± 0.57%, TCVs: –0.31% ± 0.36%;
- Training sample size n = 20: TKVs: 0.31% ± 0.51%, TCVs: –0.38% ± 0.27%.

Linear regression analysis further confirmed the strong correlation between automated and ground truth segmentations, with $R^2 > 0.999$ for both TKVs and TCVs across all four models.

Models' generalizability. **Table 3** summarizes the DSC, IoU, and HD99 segmentation performance metrics for each of the 4 ensemble models, computed on the D2 independent test set, for the kidney and cyst labels separately.

Segmentation metrics demonstrated high performance across all models. The Friedman test found no significant differences in kidney HD99 (p = 0.13); kidney DSC and IoU were borderline (p = 0.0503). For cyst segmentation, significant differences were found in DSC and IoU (χ^2 = 12.84, p = 0.005) and HD99 (p = 0.011), consistent with D1 result.

Bland–Altman analysis (see **Figure 4**) yielded the following mean relative differences (bias ± SD) between automated and ground truth segmentations:

- n5: TKVs: 0.53% ± 0.22%, TCVs: –3.07% ± 1.13%;
- n10: TKVs: 0.31% ± 0.28%, TCVs: –0.72% ± 0.96%;
- n15: TKVs: –0.23% ± 0.32%, TCVs: –0.85% ± 0.68%;
- n20: TKVs: –0.58% ± 0.47%, TCVs: –0.81% ± 1.01%.

Linear regression analysis confirmed a strong correlation, with $R^2 > 0.999$ for both TKVs and TCVs across all four models.

3.2 Automated Cyst Counting Algorithm

The algorithm worked successfully in all evaluation cases. The inference took about 1 minute on the same system used for the segmentation task (Ubuntu 22.04, 256GB of RAM, Intel Xeon 24-core, 2.50 GHz, and NVIDIA RTX A4500 GPU with 20GB of memory). The best fitness scores across the five runs ranged from 3.61% to 4.22%; the parameter set from the run achieving the lowest fitness was selected as the final configuration. The optimized parameter values, together with their sensitivity analysis results, are reported in **Supplementary Table S4** online.

Agreement and association with manual counting. The agreement between automated (hereafter referred to as Auto) and manual cyst counting on the evaluation dataset was good, with Auto bias (+3.94% vs Operator 1, and +1.39% vs Operator 2) comparable to inter-operator bias (+3.17%) and Auto limits of agreement even tighter than across operators (Auto vs Operator 1: ±10.46%, Auto vs Operator 2: ±15.81%, Operator 1 vs Operator 2: ±21.23%).

Linear regression indicated strong associations between automated and manual cyst counting. The association was highest with Operator 1 counting ($R^2 = 0.991$, $p < 0.001$) and remained very strong with Operator 2 counting ($R^2 = 0.960$, $p < 0.001$). The inter-operator correlation was the lowest ($R^2 = 0.942$, $p < 0.001$), consistent with Bland–Altman results. **Figure 5** shows both the Bland-Altman and linear regression analysis results.

Intraclass correlation coefficient analysis [ICC(2,1)] further confirmed these findings, showing excellent consistency between manual and automated cyst counting: Operator 1 vs Operator 2 (ICC = 0.940, 95% CI: 0.792-0.985), Operator 1 vs Auto (ICC = 0.982, 95% CI: 0.915-0.996), and Operator 2 vs Auto (ICC = 0.976, 95% CI: 0.905-0.994).

Performance evaluation against consensus ground truth. The three counting results (Operator 1, Operator 2, and Auto) were additionally evaluated against the consensus ground truth, defined as the average of the two operator counts. As this reference includes each operator's estimate, operator-to-consensus comparisons are not independent and should be interpreted as descriptive. The coefficient of variation (CV) was 4.09% for Auto, compared to 7.83% for Operator 1, and 7.85% for Operator 2.

Error metrics supported robust performance across methods. MAE was 4.30 cysts (95% CI: 1.90–7.00) for Auto, 5.70 (2.30–11.10) for Operator 1, and 5.50 (1.90–11.10) for Operator 2; RMSE showed a similar trend (Auto: 5.96, 2.14–8.40; Operator 1: 9.61, 2.90–15.80; Operator 2: 9.82, 2.59–16.26). Systematic bias was small for all methods (Auto: −3.50%; Operator 1: +2.70%; Operator 2: −3.30%).

Agreement with the ground truth was consistently excellent, as indicated by CCC (Auto: 0.994 [0.931–0.999], Operator 1: 0.986 [0.969–0.998], Operator 2: 0.981 [0.962–0.998]) and ICC(2,1) (ground truth vs Auto: 0.994 [0.968–0.999], ground truth vs Operator 1: 0.987 [0.951–0.997], ground truth vs Operator 2: 0.982 [0.935–0.995]; all $p < 0.001$). Linear regression further confirmed strong correlation ($R^2 = 0.992$, 0.989, and 0.981 for Auto, Operator 1, and Operator 2, respectively; all $p < 0.001$). Raw cyst counts for each evaluation case are reported in **Supplementary Table S5** online.

Sensitivity analysis. The results of the Morris screening can be found summarized as **Supplementary Table S4** and **Supplementary Fig. S4** online, highlighting the relative importance and interaction effects of the six parameters on cyst counting accuracy. Parameter σ exhibited the highest mean absolute elementary effect ($\mu^* = 2.50$), indicating the strongest overall influence on model performance. High standard deviation values for σ (1.66) and d_L (1.06) suggest notable interaction effects or non-linear responses. Most of the parameters displayed monotonic or weakly monotonic behavior, as indicated by $|\mu|/\mu^* \approx 1$.

The OAT analysis revealed that parameter σ is the only factor exerting a substantial influence on model performance when varied ±30% from its optimal value (see **Supplementary Fig. S5** online). Changes in σ produced a marked increase in MAPE beyond ±10%, confirming its dominant role. Parameters d_L, d_M, and d_S showed visible but moderate effects, with error curves remaining relatively stable compared to σ. In contrast, V_{thr} and T exhibited minimal impact, with nearly flat profiles across the tested range, indicating low sensitivity.

4 Discussion

This study introduces an operator-independent pipeline that combines an ensemble of 3D U-Net models to segment kidneys and cysts on µCT images from polycystic rat kidneys, with a morphological counting strategy optimized via a genetic algorithm to estimate the number of independent cysts from the segmentation. Our primary objective was to provide robust quantification of total kidney volume (TKV), total cyst volume (TCV), and cyst count, to streamline efficacy assessments of novel treatments in rodent models. In parallel, we investigated how the size of the training dataset affects the performance of the segmentation ensemble model and whether including a Sobel-filtered auxiliary input channel improves boundary delineation.

Deep learning performance in medical image analysis is often constrained by the size and quality of annotated training data [17]. In semantic segmentation tasks, the scarcity of annotated datasets is a major bottleneck. Cyst segmentation in polycystic kidneys is intrinsically challenging due to the strong clustering

of cysts and their potentially irregular morphology, as seen in ADPKD patients as well. These factors make the identification of the boundary between cysts and renal parenchyma prone to error, underscoring the need for robust methods that achieve high accuracy despite limited training data availability.

In segmentation, Sobel filtering is widely used for edge enhancement to refine boundary delineation [30–32]. The Sobel channel was chosen as a standard first-order edge operator: the cyst-parenchyma interface in µCT of PKD kidneys is a simple, isotropic intensity step well captured by gradient magnitude, the filter adds no trainable parameters, and it integrates into nnU-Net's multi-channel architecture without modification. Adding the Sobel channel did not yield statistically significant improvements in DSC or IoU; however, ASD and NSD analysis (Supplementary Table S3 online) showed a small, consistent improvement with the dual-channel model for cyst boundary delineation (ASD 0.017 vs. 0.019 mm; NSD 0.981 vs. 0.979 across all five test cases; dual- vs single-channel). Visual inspection supports this improvement, showing fewer boundary errors in several cases. Representative qualitative difference maps are provided as **Supplementary Fig. S2** online. Therefore, given the negligible computational overhead, the dual-channel input was retained. Alternative filtering strategies and learnable architectural components were not explored in this study; Comparative evaluation constitutes a direction for future work, especially as larger training datasets become available.Compared to Rombolotti et al. [12], the only prior study applying automatic segmentation to similar µCT images of polycystic rat kidneys, our ensemble model achieved higher performance. Their best-performing 2D U-Net variants (U-Net-bn and U-Net-ib) yielded kidney IoU ranging from 0.924 to 0.941 and Dice from 0.961 to 0.970, and cyst IoU ranging from 0.486 to 0.828 and Dice from 0.654 to 0.906 across test samples. Our 3D ensemble consistently exceeded DSC of 0.99 for kidney and 0.98 for cysts. These results are higher than those reported by Rombolotti et al. [12], but direct comparison is limited by differences in dataset composition, annotation workflow, preprocessing, train/test split, and evaluation protocol; the observed performance gap cannot be attributed solely to the choice of 2D versus 3D architecture.

For kidney segmentation, consistently high DSC and IoU across all four training-set sizes indicate comparable performance with no significant effect of dataset size. For cyst segmentation, the Friedman test identified statistically significant differences (D1: $p = 0.006$; D2: $p = 0.005$), with larger training sets tending to produce slightly higher DSC. The absolute DSC range was less than 0.003 on D1 and less than 0.009 on D2; given the small test set ($n = 5$ per dataset) and the nested, non-random nature of the training subsets, this finding should be regarded as exploratory. Taken together, these results suggest that, in this specific µCT context and under the nnU-Net framework, 15-20 annotated cases are sufficient to obtain accurate kidney and cysts masks, with kidney performance already saturating at $n = 5$. While larger datasets are typically associated with improved deep learning performance, these results support literature

demonstrating that effective models can be obtained by prioritizing high-quality annotations, robust segmentation, and well-tuned architectures [25,26]. Whether this reflects genuine robustness of the nnU-Net framework at small training set sizes or insufficient test-set diversity to reveal a stronger size effect cannot be determined from the present data; a larger, more varied evaluation set would be needed to distinguish between them.

These results are consistent with two interpretations: that the nnU-Net framework is sufficient for this task even from small training sets in this specific μCT context, or that the limited magnitude of the observed differences partly reflects insufficient test-set diversity to reveal a stronger size-dependent trend. Distinguishing the two would require a larger, more varied evaluation set.

We also evaluated the generalizability of our segmentation approach using an external test set (D2) acquired on a different μCT scanner. For translational applications, robustness across scanners and protocols is essential. While evaluation metrics decreased slightly compared with the internal validation set (D1), performance remained strong, with consistently high DSC and IoU and low HD99 values. This consistency indicates that the model tolerates data heterogeneity due to scanner differences. External validation was performed on five 14-week-old PCK rats scanned with a single additional system; validation on larger cohorts, across a broader range of ages, disease stages, acquisition systems, and multiple institutions is required to establish wider applicability.

The limited number of control SD rats included in the datasets (n = 2) reflects the primary focus of this study, which is part of a broader effort to evaluate new pharmacological treatments for ADPKD. As detailed in the Introduction, our pipeline prioritizes the quantification of TCV and cyst count over TKV, which has already been extensively characterized in the literature. The two SD control samples were included not to comprehensively represent healthy renal morphology, but rather to prevent the segmentation model from overfitting to cystic tissue alone, ensuring it could also recognize healthy parenchymal slices and thereby reduce false positives. Consistently high DSC values, exceeding 0.99 for the renal label, confirm that the inclusion of just two control samples was more than sufficient for the model to accurately delineate whole kidney, irrespective of the presence of cysts, fully serving the purposes of the present study.

Beyond volumetric biomarkers, cyst count is a direct readout of cystogenesis, rather than of their subsequent growth [14,15]. This distinction is clinically and experimentally relevant: therapies targeting cyst initiation would be expected to reduce the number of newly developed cysts before producing detectable changes in TCV or TKV, making cyst count a potentially more sensitive early endpoint for certain pharmacological interventions [10].

About cyst counting, the pipeline exhibited superior consistency compared with the manual approach, with operator-algorithm variability lower than inter-operator variability across multiple metrics, and results converging toward different operators' consensus. In particular, the Bland–Altman analysis demonstrated narrower limits of agreement and greater accuracy of the automated method with respect to manual counting. This directly addresses a key limitation of manual counting, where subjective interpretation of irregular and touching cysts can lead to substantial discrepancies between operators. Notably, even watershed-based approaches, which are often assumed to be the standard for separating touching independent objects, struggle in these cases [33,34]. The association with Operator 1 was stronger than with Operator 2 (ICC 0.982 vs 0.976; R^2 0.991 vs 0.960), consistent with the algorithm having been optimized on Operator 1 counts.

The pipeline is modular in design: the counting algorithm can operate on any binary cyst mask, whether obtained through automated segmentation, semi-automatic methods, or manual annotation. It is therefore applicable independently of the segmentation pipeline described here. If more accurate automatic segmentation methods are developed in the future, the cyst count algorithm will remain valid and useful. This task independence also applies to the experimental design: the seven samples shared between the segmentation and counting subsets do not constitute data leakage, as the two tasks involve separate training and evaluation procedures with no shared objective.

The GA-based optimization was crucial for achieving accurate cyst counting performance. Sensitivity analyses identified σ, the Gaussian smoothing applied to the DT map, as the dominant driver of accuracy, consistent with its role in suppressing spurious peaks and stabilizing center detection. Consistently, the distance-based post-processing parameters d_L, d_M, and d_S, which govern peak separation and merging in large and small cyst regions, showed moderate but non-negligible effects. In contrast, V_{thr} (the volume threshold for small/large cyst classification) and T (the multiplier used for adaptive peak thresholding) had limited impact within the explored range, suggesting that tuning these parameters alone would be unlikely to yield meaningful changes. This hierarchy of influence suggests that manual or heuristic adjustments would likely fail to capture the underlying nonlinearities and parameter interactions. The high variability observed via Morris screening also supports the need for multi-parameter optimization strategies over single-factor tuning.

Against the consensus ground truth, the automated method achieved accuracy and precision within the range of manual counting in this preliminary evaluation. ICC analysis confirmed excellent reliability (ICC > 0.95), consistent with expert-level agreement, supporting the algorithm's validity for routine use. Manual segmentation of a single µCT scan, requiring annotation across approximately 1000 slices, typically takes

several days of effort from an expert; manual cyst counting adds 1–2 hours per sample. The proposed pipeline replaces this effort with approximately 30 minutes of automated processing. This represents a significant reduction in workload and enables high-throughput analysis of large experimental cohorts.

The counting algorithm was not evaluated on D2. Unlike the segmentation model, which operates on raw μCT images and may be sensitive to scanner-specific characteristics, the counting algorithm operates exclusively on binary cyst masks. Its performance is therefore independent of the acquisition protocol. The evaluation set nonetheless challenged the algorithm across different disease stages, providing an intrinsic assessment of generalizability.

Several limitations of this study should be acknowledged. First, the ground truth segmentation masks for D1 were obtained through a semi-automatic procedure, in which a preliminary 2D U-Net provided an initial estimate that both operators then reviewed and corrected. While this introduces a dependency on the preliminary model, the correction burden was extensive, as quantified in the Methods section. The procedure also substantially reduced annotation time compared to fully manual segmentation. Second, the evaluation was performed on a limited number of cases; additional samples from different datasets, acquired from different scanners and representative of different disease stages, are needed to confirm the results. Third, the consensus ground truth for cyst counting was defined as the average of the two operator counts used for comparison; operator-to-consensus agreement analyses are therefore partly circular. Future validation of the cyst-counting algorithm should use an independent reference standard. Lastly, cyst counts reported in this study are resolution-dependent and reflect structures detectable at the 27 μm analysis resolution; thus, they should not be interpreted as an exhaustive count of all microscopic cysts potentially present at the native 9 μm acquisition resolution. This constraint aside, working at 27 μm also offers methodological advantages. At native resolution, the identification of very small structures is unreliable due to image noise, partial-volume effects, local Microfil perfusion heterogeneity, and other acquisition-related artefacts.

Future work will explore instance segmentation to enable both counting and detailed cyst-wise morphological characterization (e.g., size distributions, spatial clustering patterns). It will also address boundary-aware losses and domain adaptation strategies to further improve touching-cyst separation and cross-scanner robustness. Moreover, future studies and additional fine-tuning are needed to translate the developed models into effective tools supporting ADPKD clinical research and optimal patients' management.

5 Conclusions

To our knowledge, this work delivers the first fully integrated, operator-independent μCT pipeline combining semantic segmentation and automated cyst counting for polycystic kidney analysis in the PCK rat model. Unlike prior efforts that focused on either segmentation or counting, primarily on MRI, our integrated approach provides a complete morphological characterization toolkit tailored to the specific challenges of studying μCT scans and the polycystic condition in preclinical studies.

The operator-independent nature of the pipeline and the cross-scanner performance demonstrated on an independent external dataset provides preliminary evidence supporting its applicability across different μCT systems, while markedly reducing hands-on time and enabling high-throughput analyses, directly benefiting pharmacological research in ADPKD-oriented preclinical workflows.

Data Availability

Data used in this study are available in Zenodo (https://doi.org/10.5281/zenodo.19049543) and will be shared upon reasonable request.

Code Availability

The code-base is available on a public GitHub repository (https://github.com/Istituto-Mario-Negri-Medical-Imaging/CYSTSCAN-PKD) under an open-source license, while the model weights are stored in Zenodo (https://doi.org/10.5281/zenodo.19097241) under an open-source license. Full training hyperparameters and hardware configuration are reported in **Supplementary Table S6** to support reproducibility.

References

1. Chebib, F. T. & Torres, V. E. Autosomal Dominant Polycystic Kidney Disease: Core Curriculum 2016. *Am. J. Kidney Dis.* **67**, 792–810 (2016).

2. Walker, E. Y. X. & Marlais, M. Should we screen for intracranial aneurysms in children with autosomal dominant polycystic kidney disease? *Pediatr. Nephrol.* **38**, 77–85 (2023).

3. Torres, V. E., Rossetti, S. & Harris, P. C. Update on autosomal dominant polycystic kidney disease. *Minerva Med.* **98**, 669–691 (2007).

4. Grantham, J. J. *et al.* Volume progression in polycystic kidney disease. *N. Engl. J. Med.* **354**, 2122–2130 (2006).

5. Torres, V. E., Harris, P. C. & Pirson, Y. Autosomal dominant polycystic kidney disease. *The Lancet* **369**, 1287–1301 (2007).

6. Zhou, J. X. & Torres, V. E. Autosomal Dominant Polycystic Kidney Disease Therapies on the Horizon. *Adv. Kidney Dis. Health* **30**, 245–260 (2023).

7. Lager, D. J., Qian, Q., Bengal, R. J., Ishibashi, M. & Torres, V. E. The pck rat: A new model that resembles human autosomal dominant polycystic kidney and liver disease. *Kidney Int.* **59**, 126–136 (2001).

8. Tangri, N. *et al.* Total Kidney Volume as a Biomarker of Disease Progression in Autosomal Dominant Polycystic Kidney Disease. *Can. J. Kidney Health Dis.* **4**, 2054358117693355 (2017).

9. Grantham, J. J. & Torres, V. E. The importance of total kidney volume in evaluating progression of polycystic kidney disease. *Nat. Rev. Nephrol.* **12**, 667–677 (2016).

10. Caroli, A. & Kline, T. L. Abdominal Imaging in ADPKD: Beyond Total Kidney Volume. *J. Clin. Med.* **12**, 5133 (2023).

11. Kline, T. L. *et al.* Automatic semantic segmentation of kidney cysts in MR images of patients affected by autosomal-dominant polycystic kidney disease. *Abdom. Radiol.* **46**, 1053–1061 (2021).

12. Rombolotti, M., Sangalli, F., Cerullo, D., Remuzzi, A. & Lanzarone, E. Automatic cyst and kidney segmentation in autosomal dominant polycystic kidney disease: Comparison of U-Net based methods. *Comput. Biol. Med.* **146**, 105431 (2022).

13. Sharma, K. *et al.* Automatic Segmentation of Kidneys using Deep Learning for Total Kidney Volume Quantification in Autosomal Dominant Polycystic Kidney Disease. *Sci. Rep.* **7**, 2049 (2017).

14. Bae, K. T. *et al.* Growth Pattern of Kidney Cyst Number and Volume in Autosomal Dominant Polycystic Kidney Disease. *Clin. J. Am. Soc. Nephrol. CJASN* **14**, 823–833 (2019).

15. Bae, K. T. *et al.* Novel Methodology to Evaluate Renal Cysts in Polycystic Kidney Disease. *Am. J. Nephrol.* **39**, 210–217 (2014).

16. Bae, K. *et al.* Segmentation of Individual Renal Cysts from MR Images in Patients with Autosomal Dominant Polycystic Kidney Disease. *Clin. J. Am. Soc. Nephrol. CJASN* **8**, 1089–1097 (2013).

17. Gregory, A. V. *et al.* Semantic Instance Segmentation of Kidney Cysts in MR Images: A Fully Automated 3D Approach Developed Through Active Learning. *J. Digit. Imaging* **34**, 773–787 (2021).

18. Singh, S. P. *et al.* 3D Deep Learning on Medical Images: A Review. *Sensors* **20**, 5097 (2020).

19. Radiuk, P. Applying 3D U-Net Architecture to the Task of Multi-Organ Segmentation in Computed Tomography. *Appl Comput Syst* **25**, 43–50 (2020).

20. Ilesanmi, A. E., Ilesanmi, T. O. & Ajayi, B. O. Reviewing 3D convolutional neural network approaches for medical image segmentation. *Heliyon* **10**, (2024).

21. Yousef, R. *et al.* U-Net-Based Models towards Optimal MR Brain Image Segmentation. *Diagnostics* **13**, 1624 (2023).

22. Remuzzi, A. *et al.* Regression of Renal Disease by Angiotensin II Antagonism Is Caused by Regeneration of Kidney Vasculature. *J. Am. Soc. Nephrol. JASN* **27**, 699–705 (2016).

23. Santini, M., Fest-Santini, S. & Foltyn, P. On the local mass transfer rates around arbitrary shaped particles calculated by X-ray computed microtomography: Prospective for a novel experimental technique. *Int. Commun. Heat Mass Transf.* **79**, 135–139 (2016).

24. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

25. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).

26. Isensee, F. *et al.* nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. Preprint at https://doi.org/10.48550/arXiv.1809.10486 (2018).

27. Morris, M. D. Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics* **33**, 161–174 (1991).

28. Campolongo, F., Cariboni, J. & Saltelli, A. An effective screening design for sensitivity analysis of large models. *Environ. Model. Softw.* **22**, 1509–1518 (2007).

29. Saltelli, A. *et al.* Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environ. Model. Softw.* **114**, 29–39 (2019).

30. Sharifrazi, D. *et al.* Fusion of convolution neural network, support vector machine and Sobel filter for accurate detection of COVID-19 patients using X-ray images. *Biomed. Signal Process. Control* **68**, 102622 (2021).

31. Tang, W. *et al.* Diagnostic efficiency of multi-modal MRI based deep learning with Sobel operator in differentiating benign and malignant breast mass lesions—a retrospective study. *PeerJ Comput. Sci.* **9**, e1460 (2023).

32. Nadeem, A., Ashraf, R., Mahmood, T. & Parveen, S. Automated CAD system for early detection and classification of pancreatic cancer using deep learning model. *PloS One* **20**, e0307900 (2025).

33. Bai, M. & Urtasun, R. Deep Watershed Transform for Instance Segmentation. Preprint at https://doi.org/10.48550/arXiv.1611.08303 (2017).

34. Chen, Y., Huang, Q., Geng, M., Wang, Z. & Han, Y. A Systematic Review on Cell Nucleus Instance Segmentation. *IET Image Process.* **19**, e70129 (2025).

Funding

This work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Italian Ministry for University and Research nor the European Commission can be considered responsible for them.

Author Contributions

Andrea Mangili: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. **Alberto Arrigoni:** Conceptualization, Methodology, Software, Validation, Investigation, Writing - Review & Editing. **Fabio Sangalli**: Investigation, Resources, Data Curation, Writing - Review & Editing. **Stephanie Fest-Santini:** Resources, Data curation, Writing - Review & Editing. **Daniela Corna** Resources, Data curation, Writing - Review & Editing. **Domenico Cerullo** Resources, Data curation, Writing - Review & Editing. **Christodoulos Xinaris:** Supervision, Resources, Writing - Review & Editing. **Susanna Tomasoni:** Supervision, Resources, Writing - Review & Editing. **Maurizio Santini:** Supervision, Resources, Writing - Review & Editing. **Andrea Remuzzi:** Supervision, Writing - Review & Editing. **Anna Caroli:** Conceptualization, Validation, Supervision, Funding acquisition, Writing - Review & Editing.

Additional Information

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Figure Legends

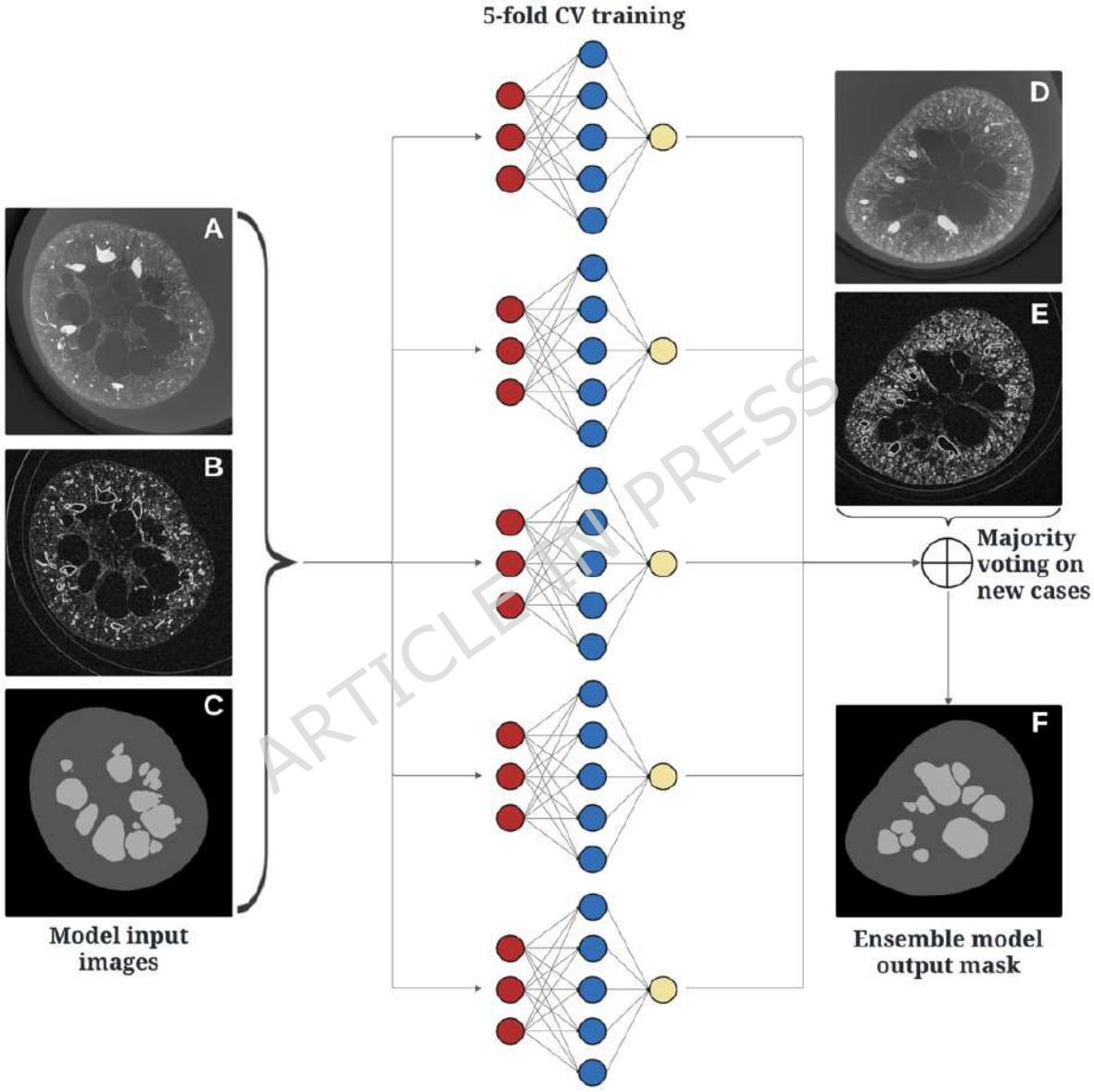


Figure 1. Pipeline of the kidney and cyst segmentation ensemble model. On the left, the dual-channel training approach is illustrated, in which both μCT images (A) and their Sobel-filtered counterparts (B) are used as input, together with the corresponding ground truth masks (C). The 5-fold cross-validation approach

involves the generation of five models. For inference, a new case is introduced to the network, with µCT images (D) and Sobel-filtered images (E) as input. An ensemble-based prediction is shown on the right, where the final segmentation (F) is obtained through a probability map averaging system among the five models. In the ground truth mask (c) and in the output mask (F), dark grey represents the kidney label and light grey represents the cyst label; black indicates background. *Abbreviations: CV = Cross Validation.*

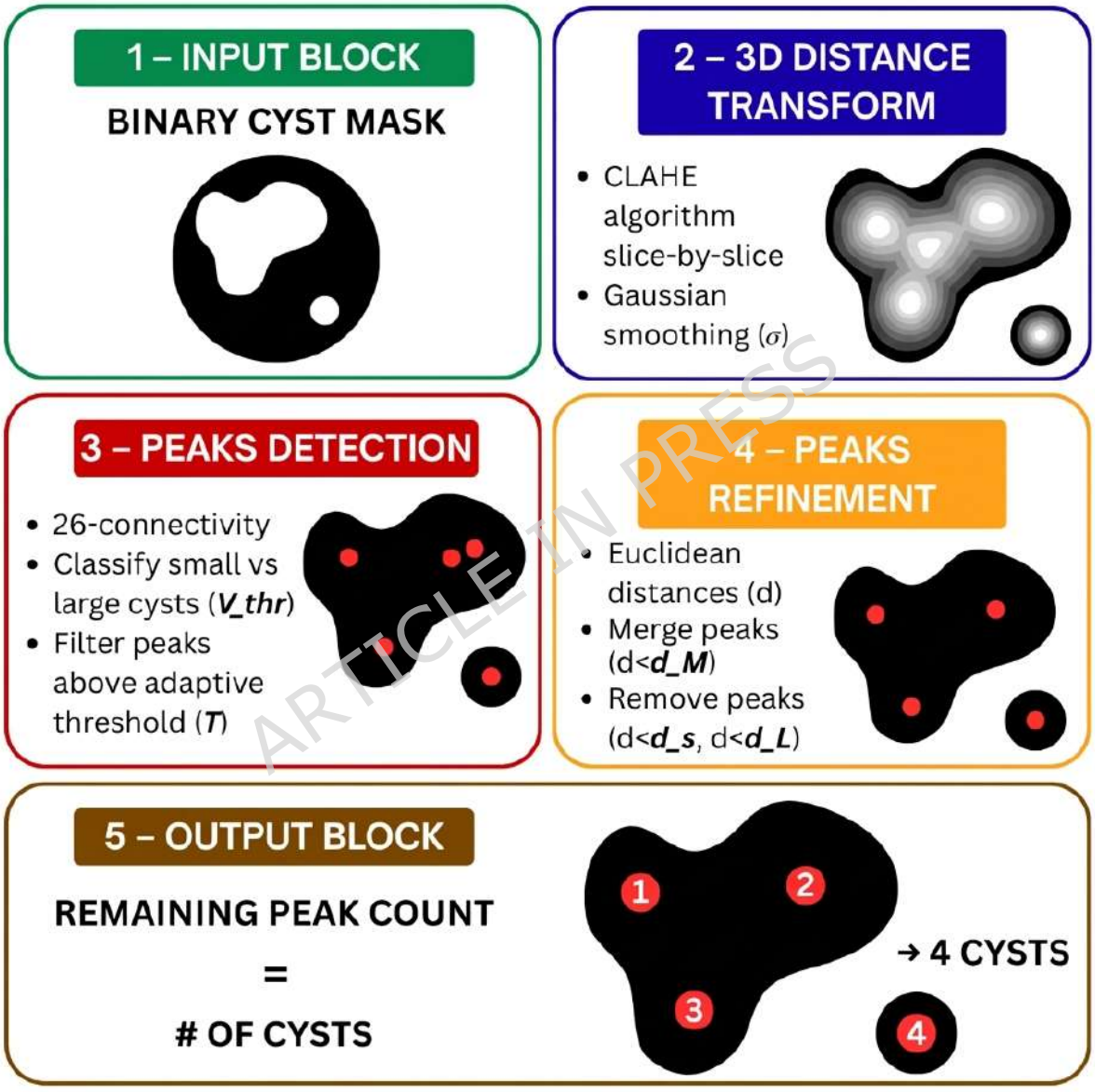


Figure 2. Workflow of the automated cyst counting algorithm. The pipeline starts from a binary cyst mask, followed by distance transform computation with Gaussian smoothing (σ). Peaks detection applies 26-connectivity analysis, cyst size classification (based on V_{thr}) and adaptive thresholding (T), while peaks

refinement merges or removes peaks based on distance criteria (d_M, d_S, d_L). The final step counts remaining peaks as individual cysts.

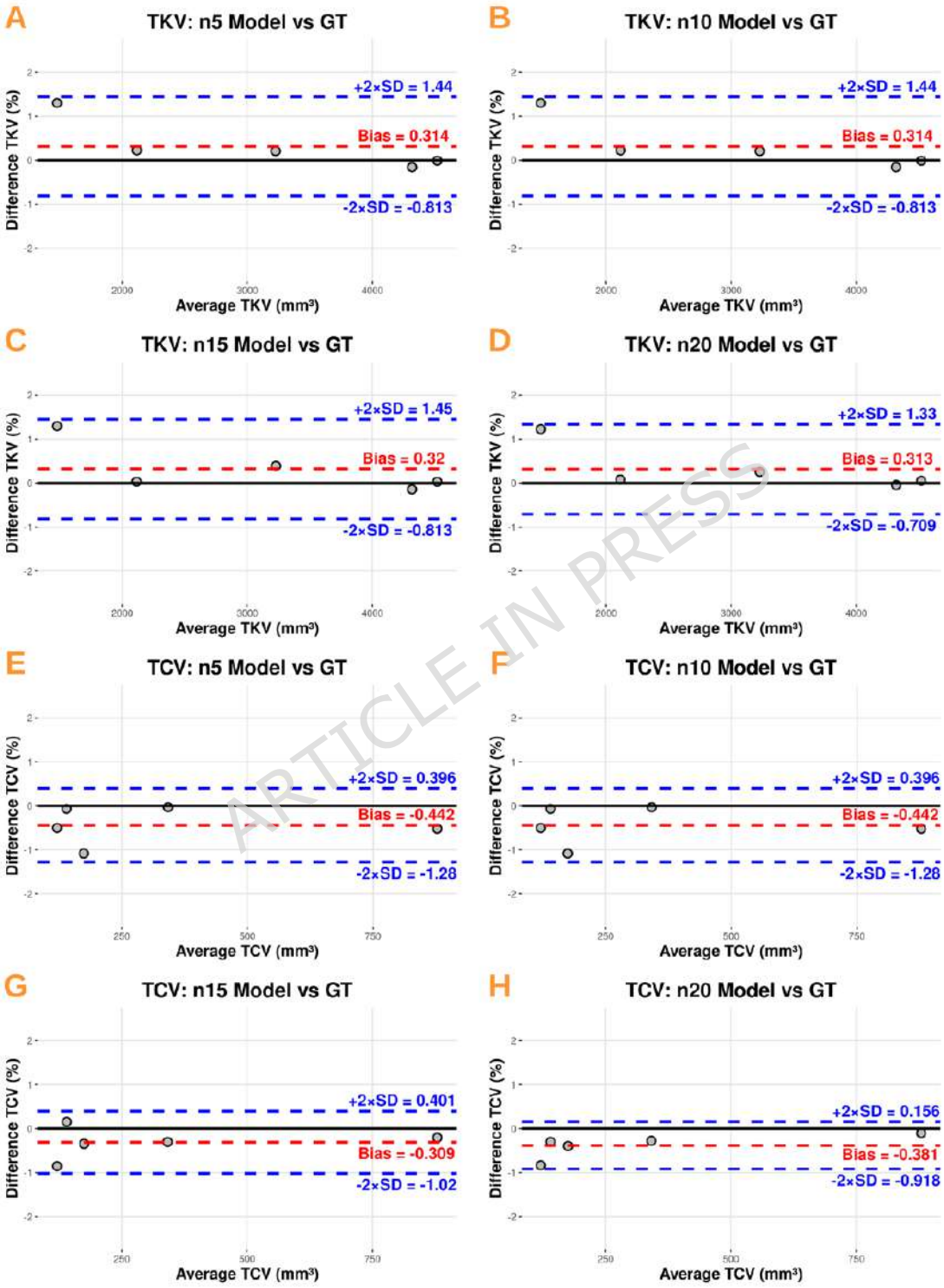


Figure 3. Kidney and cyst segmentation ensemble model performance on D1 test set. Bland-Altman plots illustrating the agreement between volumes computed from the ground truth masks and from the automatic segmentations obtained by the ensemble models trained on 5 (A: TKV, E: TCV), 10 (B: TKV, F: TCV), 15 (C: TKV, G: TCV), and 20 (D: TKV, H: TCV) cases. The results are for the D1 test set. *Abbreviations: TKV: total kidney volume; TCV: total cyst volume; GT: ground truth.*

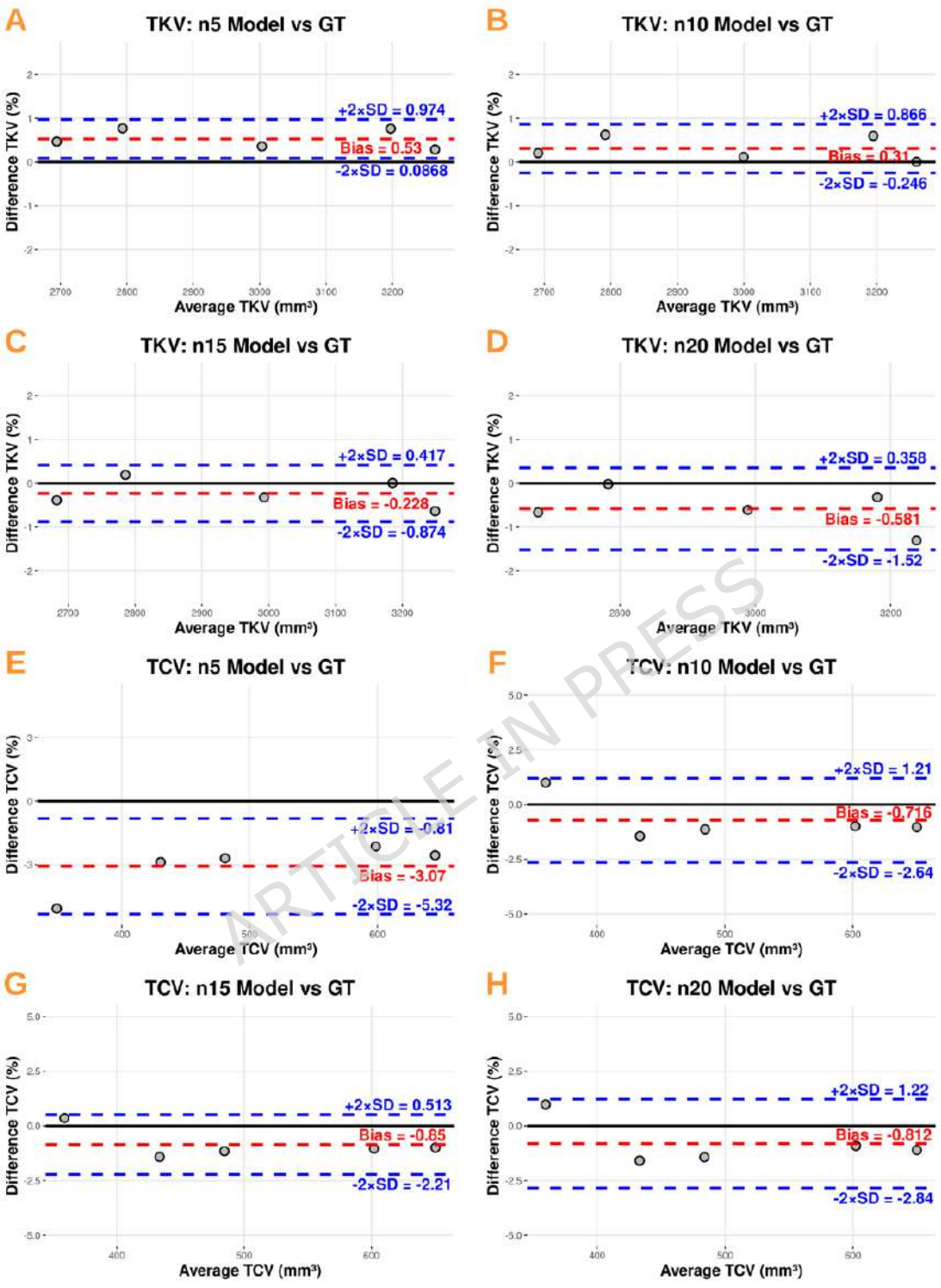


Figure 4. Kidney and cyst segmentation ensemble models generalizability on D2 test set. Bland-Altman plots illustrating the agreement between volumes computed from the ground truth masks and from the automatic segmentations obtained by the ensemble models trained on 5 (A: TKV, E: TCV), 10 (B: TKV,

F: TCV), 15 (C: TKV, G: TCV), and 20 (D: TKV, H: TCV) cases. The results are for the D2 test set. *Abbreviations: TKV: total kidney volume; TCV: total cyst volume; GT: ground truth.*

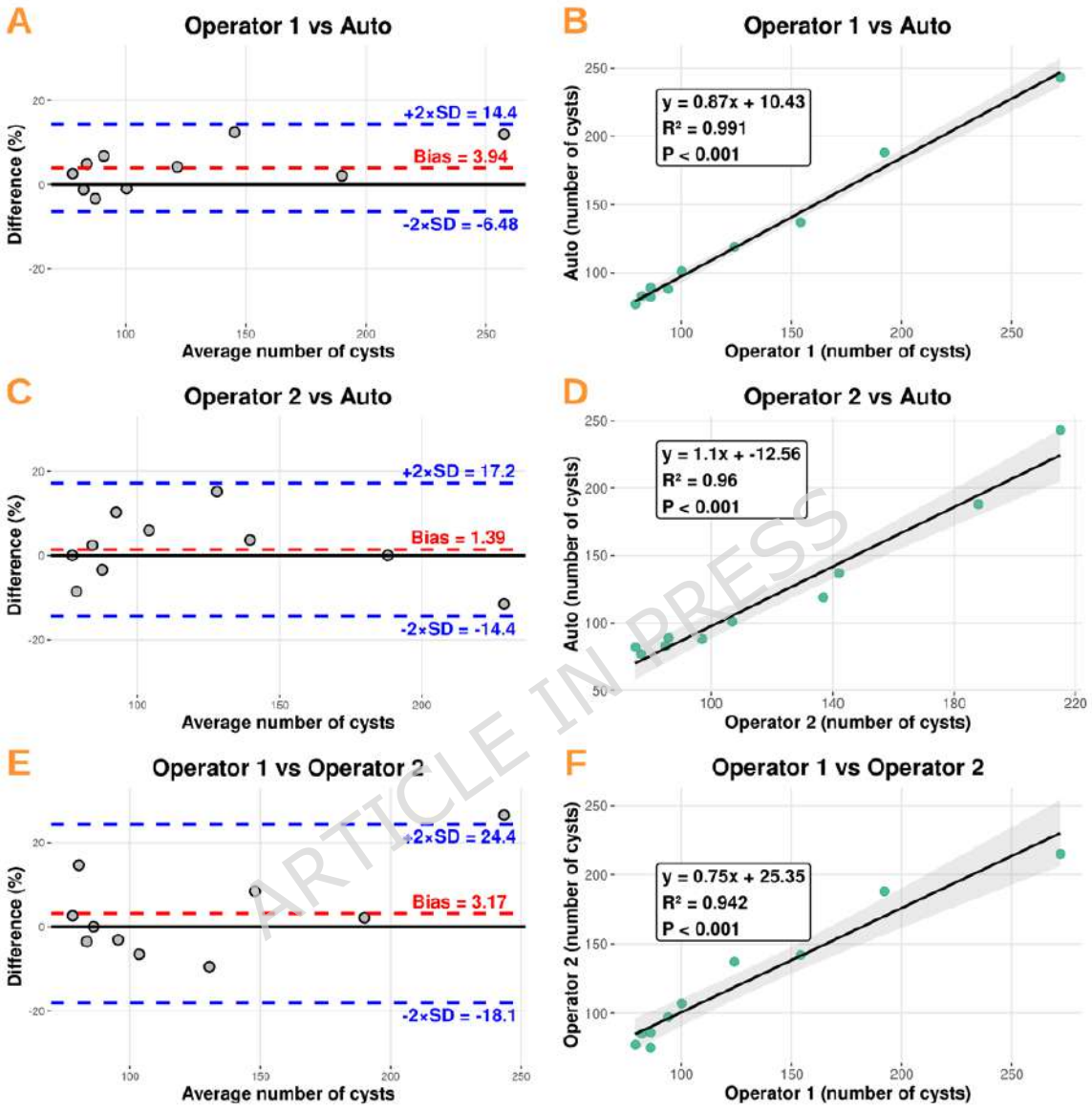


Figure 5. Automated cyst counting algorithm evaluation on 10-samples test set. Bland-Altman and linear regression analysis plots illustrating the agreement in the number of cysts between: i) the automatic algorithm and the manual Operator 1 counting (A, B), ii) the automatic algorithm and the manual Operator 2 counting (C, D), and iii) the Operator 1 and Operator 2 manual counting (E, F). The results are for the evaluation dataset.

Tables

	Dataset	Subsets	# samples	Ground truth
Kidney and cysts segmentation	*D1*	Training set	5,10,15,20	Kidney and cyst labels (O1, O2)
		Test set	5	Kidney and cyst labels (O1, O2)
	D2	External test set	5	Kidney and cyst labels (O2)
Cyst counting	*D1*	Optimization set	10	Cyst counting (O1)
		Evaluation set	10	Cyst counting (O1, O2)

Table 1. Datasets used for the development and evaluation of the automated kidney and cyst segmentation and cyst counting algorithms. This table summarizes the details of the datasets used to develop and evaluate the automated kidney and cyst segmentation ensemble model and cyst counting genetic algorithm. Seven samples were shared between the segmentation training set and the cyst counting subsets. *Abbreviations: μCT = micro-Computed Tomography; D1 = Dataset 1; D2 = Dataset 2*; *O1 = Operator 1; O2 = Operator 2.*

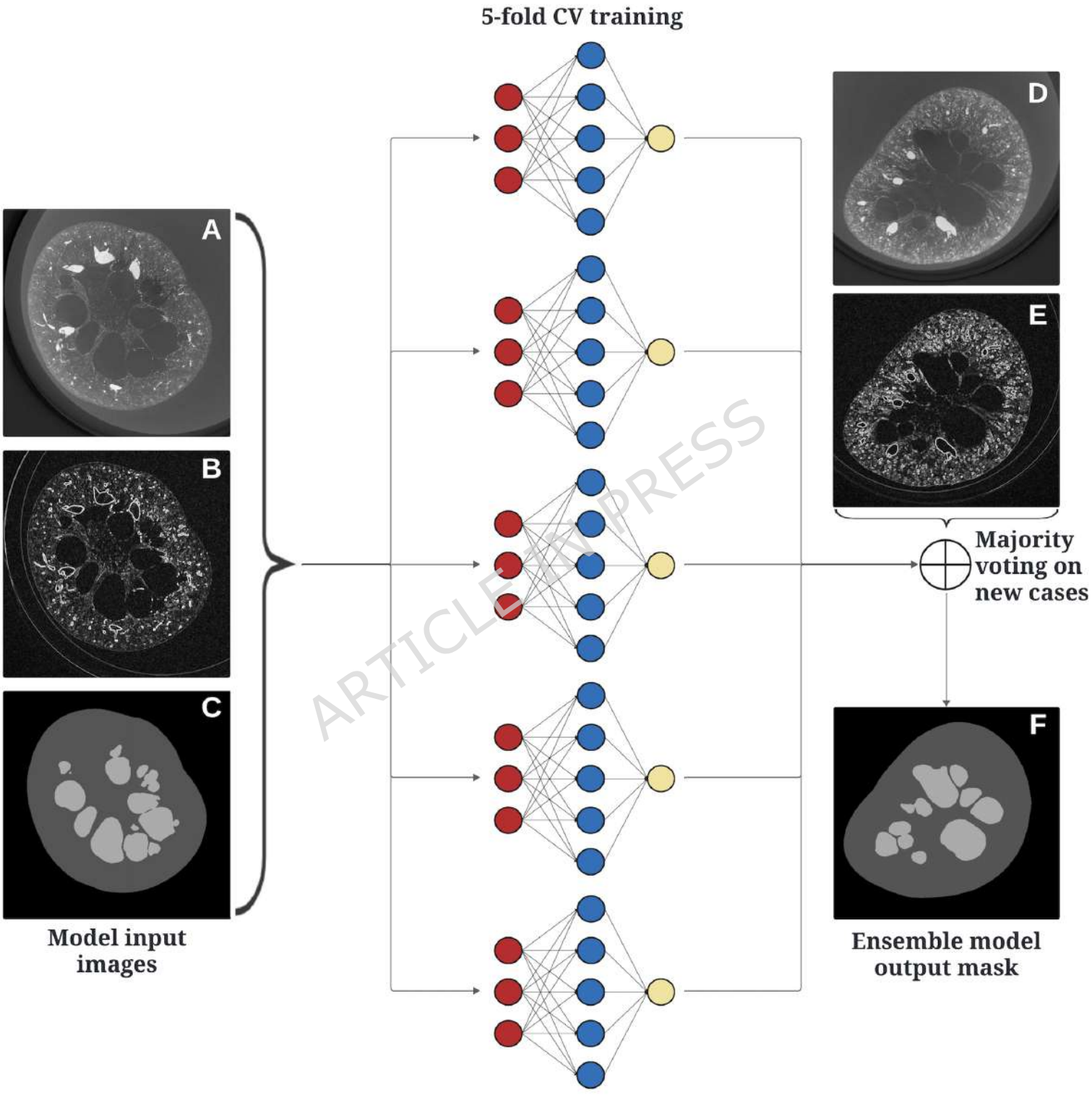
Label	Metric (mean ± SD)	Model *n5*	*n10*	*n15*	*n20*
Kidney	DSC (%)	99.54 ± 0.29	99.52± 0.36	99.57 ± 0.28	**99.61 ± 0.26**
	IoU (%)	99.10 ± 0.58	99.07 ± 0.70	99.16 ± 0.56	**99.23 ± 0.52**
	HD99 (mm)	0.40 ± 0.40	0.45 ± 0.45	0.44 ± 0.45	**0.37 ± 0.45**
Cysts	DSC (%)	98.98 ± 0.39	99.04 ± 0.48	**99.23 ± 0.34**	99.20 ± 0.39
	IoU (%)	97.98 ± 0.77	98.12 ± 0.93	**98.47 ± 0.67**	98.42 ± 0.76
	HD99 (mm)	0.31 ± 0.31	0.19 ± 0.21	**0.15 ± 0.21**	0.17 ± 0.23

Table 2. Kidney and cyst segmentation performance evaluation on the D1 test set. VDSC, IoU and HD99 metrics obtained on D1 test set by the four ensemble models trained on sets of different sizes (n5 to n20). Results are reported separately for kidney and cyst labels. The best-performing model for each metric and label is highlighted in bold. *Abbreviations: SD = standard deviation; DSC = Dice-Sorensen coefficient; IoU = Intersection over Union; HD = Hausdorff Distance 99th percentile.*

Label	Metric (mean ± SD)	Model *n5*	*n10*	*n15*	*n20*
Kidney	DSC (%)	99.04 ± 0.16	99.20± 0.13	**99.21 ± 0.15**	99.12 ± 0.27
	IoU (%)	98.10 ± 0.32	98.41 ± 0.25	**98.43 ± 0.30**	98.25 ± 0.53
	HD99 (mm)	0.37 ± 0.22	0.31 ± 0.17	0.24 ± 0.08	**0.22 ± 0.07**
Cysts	DSC (%)	97.55 ± 1.01	98.26 ± 0.56	98.28 ± 0.60	**98.43 ± 0.43**

IoU (%)	95.22 ± 1.91	96.57 ± 1.08	96.62 ± 1.15	**96.90 ± 0.83**
HD99 (mm)	0.27 ± 0.19	0.19 ± 0.12	0.18 ± 0.11	**0.16 ± 0.08**

Table 3. Kidney and cyst segmentation performance evaluation on the D2 external test set. DSC, IoU and HD99 metrics obtained on D2 test set by the four ensemble models. Results are reported separately for kidney and cysts labels. The best-performing model for each metric and label is highlighted in bold. *Abbreviations: SD = standard deviation; DSC = Dice-Sorensen coefficient; IoU = Intersection over Union; HD99 = Hausdorff Distance 99th percentile.*

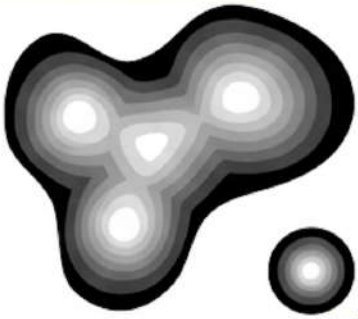


1 – INPUT BLOCK

BINARY CYST MASK

2 – 3D DISTANCE TRANSFORM

- CLAHE algorithm slice-by-slice
- Gaussian smoothing (σ)

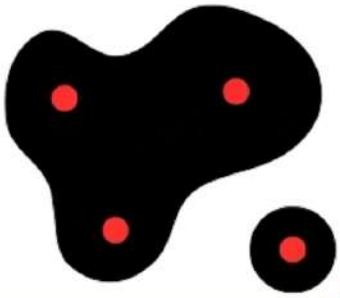


3 – PEAKS DETECTION

- 26-connectivity
- Classify small vs large cysts (***V_thr***)
- Filter peaks above adaptive threshold (***T***)

4 – PEAKS REFINEMENT

- Euclidean distances (d)
- Merge peaks (d<***d_M***)
- Remove peaks (d<***d_s***, d<***d_L***)

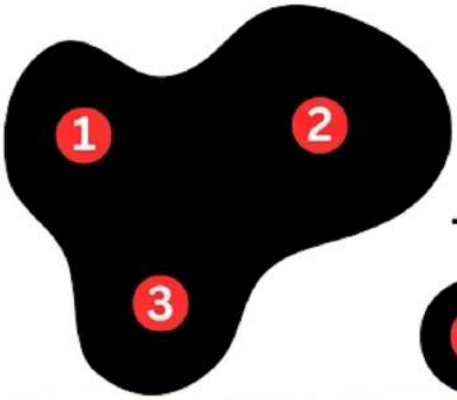


5 – OUTPUT BLOCK

REMAINING PEAK COUNT

=

 # OF CYSTS



→ 4 CYSTS