Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

UNet for the automatic segmentation of normal and pathological adrenal glands in canine CTs

Silvia Burti

silvia.burti@unipd.it

University of Padua

Artur Jurgas

AGH University of Krakow

Marek Wodzinski

Sano – Centre for Computational Personalised Medicine International Research Foundation

Andre Dourson

Antech Imaging Services

Samantha Janko

Antech Imaging Services

Nicolò Mastromattei

University of Padua

Alessandro Zotti

University of Padua

Tommaso Banzato

University of Padua

Article

Keywords:

Posted Date: August 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10379045/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

UNet for the automatic segmentation of normal and pathological adrenal glands in canine CTs

Silvia Burti[1,*], Artur Jurgas[2], Marek Wodzinski[2,3], Andre Dourson[4], Samantha Janko[4], Nicolò Mastromattei[1], Alessandro Zotti[1], and Tommaso Banzato[1]

[1]Department of Animal Medicine, Production and Health, University of Padua, Viale dell'Università 16, 35020 Legnaro, Padua, Italy
[2]AGH University of Krakow, Department of Measurement and Electronics, Krakow, PL32059, Poland
[3]Sano Centre for Computational Medicine, Krakow, PL32059, Poland
[4]Antech Imaging Services, Fountain Valley, California, USA
[*]silvia.burti@unipd.it

ABSTRACT

Computed tomography (CT) is the imaging modality of choice for the evaluation of canine adrenal glands, yet objective volumetric assessment remains limited by the need for time-consuming manual segmentation. This study developed and evaluated a deep learning–based pipeline for the automatic segmentation of normal adrenal glands and adrenal nodules using a multicentric dataset of 332 canine CT examinations acquired at more than 25 institutions. Four nnU-Net configurations were trained and compared, and performance was assessed using overlap- and distance-based metrics, volumetric agreement, and interobserver analysis. Overall, the proposed models achieved good agreement with manual annotations across all evaluated adrenal structures. The 3D low-resolution model provided the best balance between segmentation accuracy, volumetric agreement, and computational efficiency. Segmentation performance was highest for whole adrenal structures, approaching interobserver agreement, while adrenal nodules and normal adrenal glands were also segmented with good accuracy despite the greater anatomical complexity of small adrenal structures. Agreement between the model and expert readers was comparable to that reported in previous veterinary and human studies. These findings demonstrate that automated adrenal segmentation is feasible and may facilitate reproducible volumetric assessment, providing a foundation for future quantitative imaging and computer-assisted evaluation of canine adrenal disease.

Introduction

Canine adrenal gland lesions encompass a heterogeneous spectrum of conditions, ranging from benign entities, such as cortical adrenal hyperplasia and adrenal adenoma, to malignant neoplasms, including adrenocortical carcinoma and pheochromocytoma, with relevant implications for clinical management and prognosis[1,2]. Adrenal disease may be clinically suspected in dogs presenting with endocrine-related manifestations, such as hypercortisolism or systemic hypertension. However, with the increasing use of diagnostic imaging in veterinary medicine, adrenal abnormalities are increasingly detected as incidental findings, often in the absence of adrenal-specific clinical signs[3,4].

Diagnostic imaging plays a central role in the detection and characterization of adrenal gland lesions. Ultrasonography is commonly used as a first-line modality, allowing non-invasive assessment of adrenal size and morphology; however, its diagnostic performance is operator-dependent and limited by reduced sensitivity, particularly for small lesions. Although contrast-enhanced ultrasonography has been proposed to improve lesion characterization, its availability remains limited and standardized diagnostic criteria are still lacking[5]. Computed tomography (CT) is therefore widely regarded as the imaging modality of choice for adrenal gland evaluation, providing superior spatial resolution, reproducibility, and an accurate assessment of lesion size, morphology, and anatomical relationships with adjacent structures.

The increasing use of CT has resulted in more frequent detection of adrenal lesions, including incidentally discovered adrenal nodules which have been reported in up to 9.3% of dogs undergoing CT examinations for unrelated clinical indications[4]. While this growing detection rate underscores the need for objective and reproducible quantitative assessment of adrenal glands, both normal and pathological, accurate evaluation of adrenal nodules on CT remains challenging. Lesions are often small, exhibit variable enhancement patterns, and are located in proximity to major abdominal vessels and adjacent organs. In routine clinical practice, lesion characterization frequently relies on subjective assessment and linear measurements, whereas volumetric evaluation is less commonly performed due to the time-consuming nature of manual segmentation[6].

In human abdominal CT imaging, the retroperitoneal space is frequently regarded as a diagnostic "blind spot," accounting for approximately 8% of clinically significant missed findings. These misses are believed to arise primarily from perceptual

errors—due to the region's complex and variable anatomy, the uncommon occurrence of disease there, and suboptimal search patterns—rather than from any inherent limitations of the CT technology itself.[7]. In this setting, recent progress in deep learning has made it possible to create highly accurate automated methods for analysing the adrenal glands and adrenal lesions on CT in human medicine. Published models demonstrate strong detection capabilities, with sensitivities above 90% for identifying adrenal lesions, and show reliable segmentation performance, with Dice similarity coefficient (DSC) generally falling between 0.85 and 0.90, even in large and diverse datasets.[8–10]. In selected studies, model performance has been shown to be comparable to that of expert radiologists, highlighting the potential of deep learning to mitigate perceptual errors in anatomically complex regions.

In recent years, artificial intelligence has been increasingly applied in veterinary diagnostic imaging, mainly for automated lesion detection, particularly in thoracic radiography and, to a lesser extent, in musculoskeletal imaging[11–18]. AI has also been used for the classification and characterization of imaging findings, including image quality assessment and prediction of clinically relevant imaging features[14,17,19]. Deep learning–based semantic segmentation approaches have been increasingly applied to delineate anatomical structures and lesions in medical imaging[20,21], with emerging applications in veterinary CT imaging[6]. However, to date, automated tools specifically designed for canine adrenal glands are limited, with existing studies largely restricted to normal adrenal anatomy and lacking validation in pathological conditions[6]. As a result, objective and reproducible volumetric quantification of canine adrenal glands is not yet routinely available in clinical practice.

The present study seeks to design and validate an automated workflow for detecting and segmenting normal canine adrenal glands as well as adrenal nodules on CT images. Furthermore, inter-observer validation was conducted to evaluate the agreement between the automated model and expert radiologists. The proposed approach is intended to support reproducible and time-efficient lesion assessment, providing quantitative features that may be integrated into clinical reporting and research protocols.

Results

Dataset

A total of 457 CT scans were initially identified in the database. After applying the exclusion criteria, examinations were removed due to large abdominal masses displacing neighboring organs (n = 42), moderate to severe abdominal effusion (n = 23), or peritoneal metastases in contact with the target organs (n = 13), as well as motion artifacts related to respiratory movement despite general anesthesia (n = 21), suboptimal image quality (n = 17) such as poor spatial resolution preventing reliable adrenal delineation., or absence of one adrenal gland (n = 9). After these exclusions, 332 CT examinations remained eligible for inclusion. Of these, 295 scans were allocated for model training and testing using five-fold cross-validation, while 37 were set aside as a held-out test set for the assessment of inter-observer agreement.

The dogs (158 males and 174 females, mean age 8.5 years, range 2-15 years) belonged to the following breeds: American Pitbull Terrier, Australian Kelpie, Australian Shepherd, Bichon Frise, Boxer, Cavalier King Charles Spaniel, Czechoslovakian Wolfdog, Clumber Spaniel, Cockapoo, Cocker Spaniel, English Springer Spaniel, French Bulldog, German Shepherd, Golden Retriever, Great Pyrenees, Havanese, Labrador Retriever, Labrador Retriever mix, Labradoodle, Lurcher, Pinscher, Poodle, Rhodesian Ridgeback, Shar Pei, Shepherd mix, Shih Tzu mix, Siberian Husky, Staffordshire Bull Terrier, Tamaskan; the remaining dogs were mixed breed.

CT examinations were collected from more than 25 different institutions and acquired using a variety of CT scanners. The CT systems included Neurologica CereTom, Siemens Spirit, GE Medical Systems Revolution ACT, Toshiba Aquilion, Toshiba Aquilion Lightning, GE Medical Systems Brivo CT385 Series, GE Medical Systems LightSpeed Plus, Siemens Emotion 16, International Medical Solutions, GE Medical Systems LightSpeed Pro 32, GE Medical Systems LightSpeed Ultra, GE Medical Systems LightSpeed VCT, GE Medical Systems LightSpeed Plus, and GE Medical Systems LightSpeed16.

Segmentation performance and inter-observer agreement

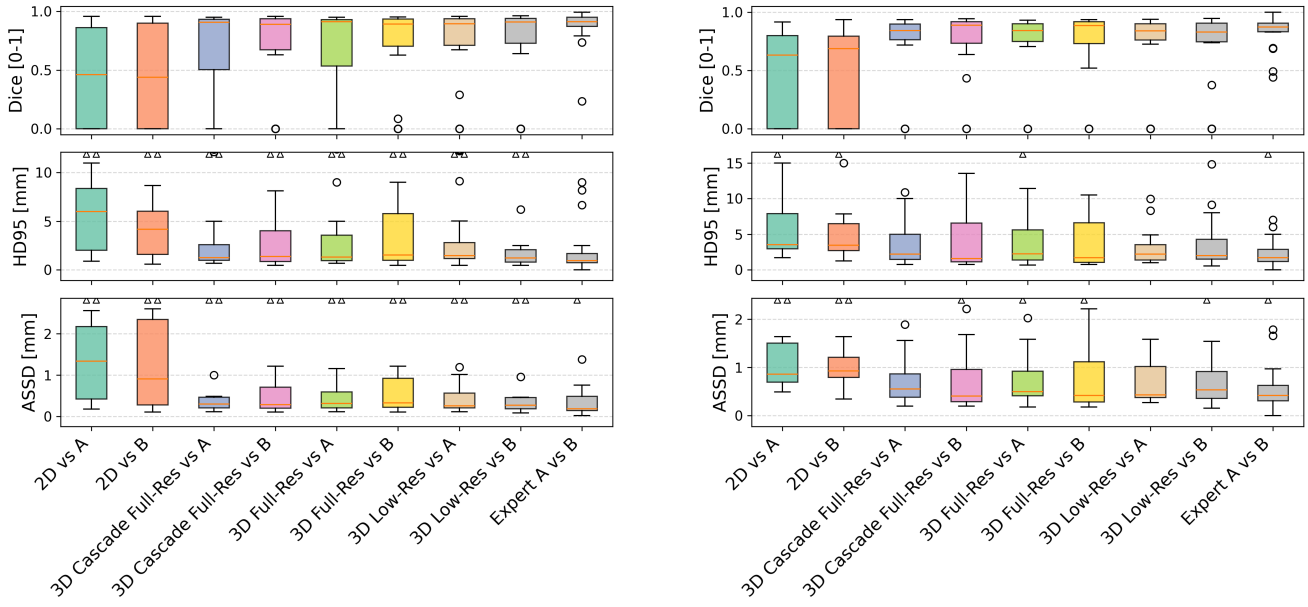
Four nnU-Net configurations—2D, 3D low-resolution, 3D full-resolution, and 3D cascade full-resolution—were trained and evaluated. Model performance was assessed on the 37-case held-out test set by comparing each model's predictions against two independent expert segmentations (Expert A and Expert B). Of the 37 test cases, 18 contained at least one adrenal nodule and 19 had only normal adrenal glands. Results are reported for three structural categories: *nodules* (the pathological parts of the adrenals), *whole adrenals* (the combined volume of the adrenal parenchyma and any associated nodule, evaluated only in subjects with nodules), and *normal adrenal glands* (left and right, evaluated only in the nodule-free cases).

After obtaining the segmentation, we applied a separate classification model that operated on the original CT image cropped to the predicted adrenal region. The classifier produced a binary output indicating the presence or absence of a nodule. If the classifier yielded a positive result, the segmentation was left unchanged. If the classifier yielded a negative result, the nodule segmentation was merged into the adrenal segmentation.

The classifier accurately recognized 65 of the 74 cases (with classification performed separately for the left and right adrenal glands), yielding an overall accuracy of 0.88. For the negative class, the precision was 0.85 and the recall was 0.92. For the positive class, the precision was 0.91 and the recall was 0.84. The left adrenal achieved an accuracy of 0.92, whereas the right adrenal achieved an accuracy of 0.84.

Pairwise comparison

Tables 1–3 summarise the pairwise segmentation metrics for each model configuration. The boxplot distributions of DSC, 95-Percentile Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD) are visualized in Fig 1. Examplary best and worst cases are shown in Fig 3.



(a) Left nodules

(b) Right nodules

Figure 1. Distribution of pairwise DSC values, HD95, and ASSD across the comparisons (model vs. Expert A, model vs. Expert B, and Expert A vs. Expert B) for left and right nodules. Some outliers were omitted for readability and are indicated by △ and ▽. Models whose distributions overlap with the expert–expert reference may be considered indistinguishable from a second human observer for this task. - DSC = Dice Similarity Score, HD = Hausdorff Distance, ASSD = Average Symmetric Surface Distance

Nodules. All 3D models achieved satisfactory nodule segmentation performance, with mean DSC approaching the inter-expert reference and, for the right adrenal, aggregated DSC values matching agreement (Tab. 1). The three 3D configurations performed comparably, with no single configuration clearly superior across both sides: the 3D low-resolution model gave the highest DSC on the left and compact distance-based error distributions, whereas the 3D full-resolution and 3D cascade full-resolution models were marginally better on the right. The cascade and full-resolution models behaved similarly to one another, with the cascade yielding somewhat lower HD95 and ASSD on the left; the full-resolution model showed the widest HD95 distributions, suggesting occasional larger localised errors. The 2D model showed the lowest overlap, with greater variability and DSC values substantially below inter-expert agreement, as well as the largest boundary errors across all distance-based metrics. Wilcoxon tests confirmed statistically significant differences between the 2D model and the expert–expert baseline ($p < 0.01$ for all comparisons). For the three 3D configurations, comparisons were largely non-significant; the only consistent exception was the left adrenal versus Expert A, which reached significance for each ($p = 0.016$–0.045), while all remaining comparisons did not differ significantly from the expert–expert baseline ($p > 0.05$; Tab. 1). As visible in the boxplots (Figure 1), while the median scores of the best-performing models approach the inter-expert reference, the interquartile ranges are somewhat wider, reflecting greater case-to-case variability than observed between the two experts.

Whole adrenals (nodule cases). Segmentation of the whole adrenal structure in cases with nodules was also satisfactory, though overall performance was somewhat lower than for nodules alone (Tab. 2). All three 3D configurations achieved mean DSCs close to the inter-expert reference, and none differed significantly from the expert–expert baseline ($p > 0.05$ in all

Table 1. Pairwise segmentation performance on **nodules** (left and right sides reported separately). Each nnUNet configuration is compared to both expert references. Wilcoxon p-values (vs Expert A vs B) are reported; **bold** p-values indicate $p > 0.05$ (not discernible from inter-rater variability); gray values indicate $p < 0.001$. Values are mean ± SD. DSC = Dice Similarity Score, HD = Hausdorff Distance, ASSD = Average Symmetric Surface Distance

Method	Side	Comparison	DSC [0-1]↑	DSC-Agg [0-1]↑	HD95 [mm]↓	ASSD [mm]↓	p-value (vs A-B)↑
Expert A	Left	vs Expert B	0.87 ± 0.17	0.945	2.07 ± 2.71	0.50 ± 0.86	-
	Right	vs Expert B	0.82 ± 0.15	0.858	3.72 ± 6.38	0.98 ± 1.81	-
2D	Left	vs Expert A	0.45 ± 0.39	0.310	8.25 ± 9.43	2.71 ± 4.41	<0.001
		vs Expert B	0.46 ± 0.41	0.310	7.06 ± 9.54	2.65 ± 4.57	0.002
	Right	vs Expert A	0.49 ± 0.38	0.461	6.03 ± 5.18	1.38 ± 1.13	<0.001
		vs Expert B	0.50 ± 0.38	0.456	5.73 ± 5.46	1.38 ± 1.24	<0.001
3D Low-Res	Left	vs Expert A	0.72 ± 0.35	0.763	3.38 ± 4.08	0.87 ± 1.42	0.045
		vs Expert B	0.74 ± 0.34	0.761	3.33 ± 5.20	0.77 ± 1.29	**0.418**
	Right	vs Expert A	0.71 ± 0.33	0.848	3.25 ± 2.65	0.69 ± 0.45	**0.284**
		vs Expert B	0.69 ± 0.34	0.810	3.87 ± 3.94	0.91 ± 1.23	**0.108**
3D Full-Res	Left	vs Expert A	0.68 ± 0.38	0.467	4.47 ± 7.32	1.53 ± 3.61	0.029
		vs Expert B	0.70 ± 0.37	0.469	5.36 ± 8.17	1.56 ± 3.68	**0.104**
	Right	vs Expert A	0.75 ± 0.28	0.873	4.61 ± 5.45	0.83 ± 0.75	**0.099**
		vs Expert B	0.74 ± 0.29	0.858	3.89 ± 3.68	0.81 ± 0.80	**0.229**
3D Cascade Full-Res	Left	vs Expert A	0.67 ± 0.39	0.792	3.02 ± 3.92	0.84 ± 1.39	0.016
		vs Expert B	0.69 ± 0.38	0.789	3.95 ± 5.37	0.84 ± 1.24	**0.113**
	Right	vs Expert A	0.75 ± 0.28	0.873	3.64 ± 3.31	0.73 ± 0.52	**0.119**
		vs Expert B	0.74 ± 0.30	0.858	4.01 ± 4.16	0.89 ± 1.09	**0.196**

comparisons). The 3D full-resolution model produced the most contained distance-based errors overall, with the 3D cascade full-resolution model matching it for the right adrenal. The 3D low-resolution model, despite comparable DSC, showed markedly inflated HD95 and ASSD with very large standard deviations on both sides, indicating occasional large-outlier cases, while the cascade model exhibited the largest left-sided distance errors, likewise driven by outliers. The 2D model again showed the lowest overlap, with statistically significant differences from the expert–expert baseline on both sides ($p < 0.05$). As with nodules, boxplot distributions for the 3D models largely overlapped with the expert–expert reference, though with greater spread and occasional low-scoring outliers.

Normal adrenal glands. Segmentation of normal adrenal glands (in the nodule-free subset) was more variable than the other categories, particularly for the right adrenal where inter-expert agreement was itself lower (Tab. 3). Most setups yielded strong DSC values that came close to the level of agreement seen between human experts. The 3D full-resolution model was the strongest, matching or slightly exceeding the inter-expert DSC on the left with very compact distance errors (HD95 ≈1.2–1.3 mm). Critically, none of the three 3D configurations differed significantly from the expert–expert baseline on either side or against either expert ($p > 0.05$ throughout), indicating that their normal-adrenal segmentations were not statistically discernible from inter-rater variability; the 2D model was the only configuration to reach significance, and in a single comparison only (right adrenal versus Expert B, $p = 0.005$). One notable feature remains a dissociation between overlap and distance metrics: despite their high DSC, the 3D low-resolution and 3D cascade full-resolution models produced HD95 and ASSD values inflated to very large magnitudes with extremely high standard deviations (e.g. left-sided HD95 of 34.7 ± 97.4 and 55.5 ± 125.2 mm, respectively). This is driven by a small number of outlier cases rather than a systematic boundary error, and most likely reflects the sensitivity of distance-based metrics to occasional spurious or displaced predicted components when the target structure is small. The 3D full-resolution and 2D models, by contrast, combined high overlap with well-behaved distance metrics.

Volumetric agreement

Bland–Altman plots (Figure 2) were used to assess volumetric agreement between each model and the mean of the two expert segmentations. Differences were computed as the model-predicted volume minus the mean expert volume; a negative mean bias therefore indicates that the model systematically *under-segments* (predicts smaller volumes than the experts), whereas a positive mean bias indicates systematic *over-segmentation*. The solid line denotes the mean bias and the dashed lines the 95% limits of agreement (LoA), with narrower limits reflecting more consistent agreement and a trend of the differences with mean volume reflecting a proportional, volume-dependent bias.

Table 2. Pairwise segmentation performance on the **combined volume of the nodules and the radiographically normal portion of the adrenal (if present)** (left and right sides reported separately). Each nnUNet configuration is compared to both expert references. Wilcoxon p-values (vs Expert A vs B) are reported; **bold** p-values indicate $p > 0.05$ (not discernible from inter-rater variability); gray values indicate $p < 0.001$. Values are mean ± SD. DSC = Dice Similarity Score, HD = Hausdorff Distance, ASSD = Average Symmetric Surface Distance

Method	Side	Comparison	DSC [0-1]↑	DSC-Agg [0-1]↑	HD95 [mm]↓	ASSD [mm]↓	p-value (vs A-B)↑
Expert A	Left	vs Expert B	0.89 ± 0.06	0.944	0.99 ± 0.44	0.23 ± 0.12	-
	Right	vs Expert B	0.80 ± 0.22	0.867	4.31 ± 10.02	2.15 ± 6.90	-
2D	Left	vs Expert A	0.77 ± 0.25	0.460	3.82 ± 6.81	1.43 ± 2.96	0.005
		vs Expert B	0.77 ± 0.25	0.461	3.94 ± 6.71	1.47 ± 3.01	0.001
	Right	vs Expert A	0.76 ± 0.23	0.747	3.99 ± 5.54	1.27 ± 2.37	0.035
		vs Expert B	0.70 ± 0.29	0.737	5.97 ± 9.46	2.83 ± 6.98	0.011
3D Low-Res	Left	vs Expert A	0.87 ± 0.11	0.780	17.73 ± 69.05	5.86 ± 26.10	**0.633**
		vs Expert B	0.87 ± 0.12	0.778	17.71 ± 69.03	5.66 ± 25.33	**0.988**
	Right	vs Expert A	0.86 ± 0.07	0.894	14.85 ± 79.95	1.66 ± 7.77	**0.068**
		vs Expert B	0.79 ± 0.22	0.870	17.20 ± 80.28	3.43 ± 10.26	**0.844**
3D Full-Res	Left	vs Expert A	0.86 ± 0.13	0.539	4.68 ± 17.14	0.88 ± 2.58	**0.428**
		vs Expert B	0.87 ± 0.13	0.545	4.64 ± 16.96	0.85 ± 2.59	**0.788**
	Right	vs Expert A	0.85 ± 0.08	0.887	2.15 ± 3.40	0.44 ± 0.50	**0.187**
		vs Expert B	0.80 ± 0.22	0.876	4.07 ± 9.23	2.11 ± 7.04	**0.149**
3D Cascade Full-Res	Left	vs Expert A	0.86 ± 0.12	0.805	27.87 ± 90.26	8.06 ± 28.96	**0.192**
		vs Expert B	0.87 ± 0.12	0.804	27.87 ± 90.27	7.83 ± 28.17	**0.591**
	Right	vs Expert A	0.85 ± 0.07	0.888	1.71 ± 1.36	0.49 ± 0.61	**0.225**
		vs Expert B	0.80 ± 0.22	0.875	4.05 ± 9.30	2.24 ± 7.06	**0.154**

Table 3. Pairwise segmentation performance on **normal adrenals** (left and right sides reported separately). Each nnUNet configuration is compared to both expert references. Wilcoxon p-values (vs Expert A vs B) are reported; **bold** p-values indicate $p > 0.05$ (not discernible from inter-rater variability); gray values indicate $p < 0.001$. Values are mean ± SD. - DSC = Dice Similarity Score, HD = Hausdorff Distance, ASSD = Average Symmetric Surface Distance

Method	Side	Comparison	DSC [0-1]↑	DSC-Agg [0-1]↑	HD95 [mm]↓	ASSD [mm]↓	p-value (vs A-B)↑
Expert A	Left	vs Expert B	0.87 ± 0.05	0.879	1.08 ± 0.43	0.25 ± 0.11	-
	Right	vs Expert B	0.73 ± 0.28	0.844	5.80 ± 12.56	3.49 ± 9.40	-
2D	Left	vs Expert A	0.81 ± 0.18	0.855	2.42 ± 4.79	0.75 ± 1.51	**0.766**
		vs Expert B	0.81 ± 0.18	0.854	2.54 ± 4.78	0.77 ± 1.54	**0.393**
	Right	vs Expert A	0.70 ± 0.23	0.779	3.66 ± 3.58	0.98 ± 1.00	**0.123**
		vs Expert B	0.58 ± 0.32	0.725	8.02 ± 11.78	4.28 ± 9.34	0.005
3D Low-Res	Left	vs Expert A	0.83 ± 0.14	0.842	34.74 ± 97.43	11.61 ± 37.06	**0.899**
		vs Expert B	0.84 ± 0.14	0.856	34.70 ± 97.40	11.20 ± 35.98	**0.702**
	Right	vs Expert A	0.78 ± 0.16	0.830	27.48 ± 111.52	2.89 ± 10.84	**0.651**
		vs Expert B	0.68 ± 0.29	0.794	31.69 ± 111.48	6.18 ± 13.89	**0.332**
3D Full-Res	Left	vs Expert A	0.86 ± 0.05	0.862	1.28 ± 1.41	0.30 ± 0.25	**0.865**
		vs Expert B	0.88 ± 0.04	0.882	1.23 ± 1.38	0.27 ± 0.24	**0.212**
	Right	vs Expert A	0.71 ± 0.31	0.798	2.05 ± 1.62	0.59 ± 0.64	**0.768**
		vs Expert B	0.61 ± 0.37	0.762	6.44 ± 12.43	3.88 ± 9.61	**0.492**
3D Cascade Full-Res	Left	vs Expert A	0.81 ± 0.15	0.822	55.48 ± 125.20	16.04 ± 40.54	**0.393**
		vs Expert B	0.83 ± 0.15	0.836	55.47 ± 125.21	15.58 ± 39.44	**0.766**
	Right	vs Expert A	0.71 ± 0.31	0.805	1.95 ± 1.48	0.73 ± 0.93	**0.651**
		vs Expert B	0.61 ± 0.37	0.764	6.23 ± 12.43	4.01 ± 9.61	**0.492**

For nodules, the best-performing configuration differed by side, mirroring the ICC results below. For the right adrenal, the 3D full-resolution and 3D cascade full-resolution models showed essentially no mean bias (-0.02 and -0.07 mL) with the narrowest limits of agreement; the 3D low-resolution model was intermediate (-0.55 mL), and the 2D model exhibited the largest negative bias (-1.74 mL) together with the widest limits and a prominent under-segmentation outlier. For the left adrenal, where lesion volumes spanned a much wider range, the 3D cascade full-resolution and 3D low-resolution models gave the smallest biases (-1.41 and -1.80 mL) and clearly the narrowest limits of agreement, whereas the 3D full-resolution and 2D models under-segmented more strongly (-3.44 and -5.46 mL) with substantially wider limits; the single most extreme error was a large left-sided lesion grossly under-segmented by the 2D model. Across both sides, the 2D model consistently showed the largest negative bias and widest limits, indicating systematic under-segmentation of nodule volume.

For normal adrenal glands the pattern again depended on side and configuration. On the right, all four configurations achieved good volumetric agreement, with small negative biases (-0.07 to -0.17 mL) and comparable limits of agreement; the 3D full-resolution model had the smallest bias and the 2D model the largest. On the left, the 3D full-resolution and 2D models were highly accurate, with near-zero bias and tight limits of agreement, whereas the 3D low-resolution and 3D cascade full-resolution models showed a positive bias ($+0.27$ and $+0.39$ mL) and markedly wider limits, driven by occasional over-segmentation of the normal gland (including one clear high outlier). This residual tendency of the low-resolution and cascade configurations to over-segment the left normal adrenal is consistent with their lower volumetric ICC for this structure (below) and with the occasional large distance-metric outliers reported earlier.

Intraclass correlation coefficients

ICC values for volumetric measurements among Expert A, Expert B, and the model are reported in Tab. 4. Because both experts contribute to every estimate, differences in ICC across model configurations can be read as comparisons of model–expert agreement on a common scale anchored to inter-expert variability.

For nodule volumes, the best configuration depended on laterality. For left nodules, the 3D cascade full-resolution model achieved the highest ICC (0.939), marginally above 3D low-resolution (0.910), while the 2D (0.507) and 3D full-resolution (0.580) models were substantially lower. For right nodules, the 3D full-resolution model led (0.838) over the closely grouped cascade (0.824) and low-resolution (0.807) models, with the 2D model again the lowest (0.593). Overall, the 3D low-resolution and 3D cascade full-resolution models provided the most consistent high agreement for nodule volumetry across both sides, whereas the 3D full-resolution model was strong on the right but weak on the left.

For whole adrenal volumes in nodule cases, agreement was again side-dependent. On the left, the 3D cascade full-resolution (0.944) and 3D low-resolution (0.916) models showed excellent agreement, whereas the 3D full-resolution (0.630) and 2D (0.546) models were considerably weaker. On the right, all three 3D configurations were closely grouped and high (0.868–0.886, led marginally by the full-resolution model) and clearly outperformed the 2D model (0.761). The cascade and low-resolution configurations were therefore the most reliable for left-sided whole-adrenal volumetry, while all three 3D configurations performed comparably well on the right.

For normal adrenal glands, the earlier impression of a collapse of agreement with increasing model complexity was not borne out. On the right, all four configurations achieved high ICC values (0.867–0.926, led by the 3D low-resolution model), indicating uniformly strong volumetric agreement. On the left, the 3D full-resolution model gave the highest ICC (0.909), with the 2D model also performing well (0.857), whereas the 3D low-resolution (0.496) and 3D cascade full-resolution (0.464) models dropped substantially. This left-sided weakness of the low-resolution and cascade configurations is consistent with the positive bias and wider limits of agreement seen for them in the Bland–Altman analysis and points to occasional over-segmentation of the normal gland rather than a systematic failure.

Table 4. Intraclass Correlation Coefficients (ICC) for volumetric measurements among the three observers (Expert A, Expert B, and the model), stratified by structure group and nnU-Net configuration. ICC values were computed using a two-way random-effects model with absolute agreement; bolded values indicate the highest ICC in each row.

Structure group	2D	3D Low-Res	3D Full-Res	3D Cascade Full-Res
Whole left adrenals	0.546	0.916	0.630	**0.944**
Whole right adrenals	0.761	0.868	**0.886**	0.874
Left nodules	0.507	0.910	0.580	**0.939**
Right nodules	0.593	0.807	**0.838**	0.824
Normal left adrenals	0.857	0.496	**0.909**	0.464
Normal right adrenals	0.867	**0.926**	0.902	0.907

Inference time

Mean inference times per case are reported in Tab. 5. The 2D model was the fastest and the 3D cascade full-resolution model the most computationally demanding, with the 3D full-resolution model intermediate and 3D low-resolution considerably faster than either full-resolution variant. Taken together with the segmentation results, the 3D low-resolution model offers the most favourable overall balance: it achieves the best or near-best nodule segmentation performance, competitive whole adrenal performance, and avoids the severe degradation on normal adrenals shared by the two full-resolution architectures—all while being considerably less computationally expensive than either. The 3D cascade full-resolution model, despite achieving the highest ICC for left-sided nodules and left adrenals, did not offer a consistent advantage over the simpler 3D full-resolution model and introduced the highest computational cost, making it difficult to justify as a preferred configuration.

Table 5. Average inference and train time per case for each model configuration. Timings for the classification model are reported separately, rather than being added to the segmentation configurations.

Model	Mean inference time [s]↓	Mean data processing time [s]↓	Training time [h]↓
Classification	0.21	2.45	44
2D	94.62	5.20	28
3D Low-Resolution	564.01	3.43	33
3D Full-Resolution	1539.14	3.67	42
3D Cascading Full-Resolution	2103.15	6.55	115

Discussion

Overall segmentation metrics indicated good agreement between automated predictions and manual annotations across the different adrenal structures evaluated. The highest performance was observed for whole adrenal structures, with automated segmentations approaching the level of agreement observed between human observers. Slightly lower but still favourable results were obtained for adrenal nodules and normal adrenal glands, indicating that the proposed models were able to reliably identify both pathological and non-pathological adrenal tissue. Distance-based metrics further supported these findings, showing generally close correspondence between predicted and reference contours across all evaluated structures. Among the evaluated architectures, the 3D models consistently outperformed the 2D configuration, with the 3D low-resolution model providing the most favourable balance between segmentation accuracy, volumetric agreement, and computational efficiency.

One potential limitation affecting the reproducibility of adrenal nodule segmentation is the frequent lack of a clearly definable boundary between normal adrenal parenchyma and pathological nodular tissue. In many cases, adrenal nodules do not present as sharply demarcated masses, but rather as focal or multifocal enlargements that gradually merge with the remaining gland. This is particularly relevant for small nodules, mildly enlarged adrenal glands, or lesions with attenuation and contrast enhancement characteristics similar to those of the adjacent normal adrenal tissue. Consequently, the decision on where the nodule ends and the non-nodular portion of the gland begins may be partly subjective and depends on the interpretation of the individual operator. This ambiguity affects not only automated segmentation but also manual annotation, and may therefore contribute to the interobserver variability observed in this study.

For adrenal nodules, all three 3D architectures achieved mean DSC values approaching interobserver agreement, with mean DSC ranging from 0.67 to 0.75 compared with 0.82–0.87 between the two experts (Tab. 1). Although overlap remained generally good, distance-based metrics demonstrated greater variability than observed for expert–expert agreement. The best-performing models achieved mean ASSD values below 1 mm for most comparisons, whereas HD95 values generally ranged between approximately 3 and 5 mm, compared with 2.07 mm and 3.72 mm for the expert–expert reference on the left and right adrenal glands, respectively (Tab. 1). These findings indicate that the models were generally able to localize adrenal nodules reliably, while accurate delineation of lesion margins remained more challenging in morphologically complex cases. This behaviour likely reflects the marked heterogeneity of adrenal nodules, which differ substantially in morphology, enhancement pattern, and relationship with adjacent anatomical structures. Performance further improved when the entire adrenal gland, including both normal parenchyma and nodular tissue, was considered as a single structure. Most whole adrenal segmentations achieved mean DSC values of up to 0.87 and ASSD values below 1 mm (Tab. 2), indicating excellent overall agreement with manual segmentations. This improvement is expected because defining the external adrenal contour is generally less challenging than accurately identifying the interface between normal and nodular adrenal tissues. From a clinical perspective, whole-adrenal segmentation may represent the most relevant target for applications such as automated adrenal volumetry, longitudinal monitoring, surgical planning, and treatment response assessment, where precise subdivision between normal and abnormal tissue is often less critical than accurate delineation of the gland as a whole. Normal adrenal glands also

showed good segmentation performance, although mean DSC values remained slightly lower than those obtained for whole adrenal segmentation. This finding is not unexpected, as the adrenal glands represent small anatomical structures in which even minimal contour discrepancies can produce substantial reductions in overlap-based metrics. Importantly, boundary-based metrics remained consistently low for the best models, with ASSD values around 0.5–0.7 mm, indicating that predicted and manual contours remained closely aligned despite modest reductions in DSC. The left adrenal gland consistently achieved slightly better segmentation performance than the right, a finding that is consistent with the known anatomical characteristics of the canine adrenal glands. In particular, the right adrenal gland is closely associated with the liver, making precise delineation of its boundaries inherently more challenging. This apparent discrepancy between overlap- and distance-based metrics further highlights the importance of interpreting multiple complementary evaluation measures when assessing segmentation performance for very small anatomical structures. Overall, overlap- and distance-based metrics remained generally favorable, indicating that the proposed approach is capable of reliably identifying normal adrenal anatomy despite the small size of the target structure. Interobserver agreement provided an essential clinical benchmark for interpreting model performance. Interestingly, the interobserver analysis showed a higher overall level of agreement among readers for pathological cases than for normal ones, particularly for the right adrenal. For the left adrenal, agreement between observers was almost the same whether they assessed nodules alone, nodules plus the radiographically normal adrenal tissue, or the entire adrenal gland. In contrast, for the right adrenal, interobserver agreement differed markedly between nodular and normal adrenal tissue. One possible explanation is that some subtle differences in segmentation strategy may occur between the observers. For example, in some cases it is challenging to reliably exclude the phrenico abdominal artery from the segmentation, while some operators may tend to slightly under-segment or over-segment the gland. Because the adrenals are very small structures, even minor discrepancies between segmentations can lead to substantial reductions in the DSC values.

Automatic segmentation of canine adrenal glands has previously been described in only one study[6], which considered solely normal adrenal glands. In[6], the authors employed a multicentric design, but neither the number of centers nor their geographic locations were specified. That work reported a DSC of 0.885 ± 0.075, and no distance-based metrics were provided. In the present study, the DSC values for the left adrenal gland are comparable to those reported by[6], whereas the DSC for the right adrenal gland is slightly lower. We propose that one contributing explanation is the substantially larger number of centers participating in our study, along with the increased heterogeneity in overall image quality, which may account for the observed discrepancy. Another important distinction is that[6] excluded cases in which the interface between the right adrenal gland and the liver was poorly delineated. By contrast, we intentionally included all cases in which the adrenal glands were visible, to better approximate real-world clinical conditions.

Furthermore, the agreement achieved between the proposed model and the expert readers was comparable to that reported by Robinson-Weiss et al.[22] for human adrenal CT segmentation. In that study, the authors reported mean DSC values of approximately 0.80 for normal adrenal glands and 0.84 for adrenal masses, values that closely match those observed in the present study. More broadly, recent studies in human medicine have expanded the application of deep learning beyond segmentation to include automated lesion characterization and risk stratification. However, direct comparison with veterinary medicine remains challenging, as standardized CT-based criteria for differentiating benign from malignant adrenal lesions have not yet been established in dogs. Interestingly, our model was able to accurately identify and segment adrenal nodules within the adrenal gland, representing a potentially relevant advancement compared with previous reports in human medicine, where automated segmentation of individual nodules in nodular adrenal hyperplasia has remained challenging[23]. Overall, these comparisons suggest that the proposed approach achieves performance comparable to that reported in both the veterinary and human literature, despite being evaluated on a more heterogeneous and clinically representative dataset. Taken together, these findings indicate that the proposed framework performs at a level comparable with current veterinary and human studies, despite being evaluated on a substantially more heterogeneous and clinically representative dataset.

The segmentation performance observed in the present study suggests that automated adrenal segmentation may have several clinically relevant applications. In routine veterinary practice, CT assessment of adrenal lesions is still largely based on subjective interpretation and linear measurements, whereas volumetric analysis is rarely performed because manual segmentation is time-consuming and poorly reproducible. The proposed approach enables rapid and standardized delineation of adrenal structures, providing objective volumetric measurements that may facilitate lesion characterization, treatment planning, and longitudinal follow-up. Beyond routine clinical evaluation, automated segmentation could also facilitate the extraction of quantitative imaging biomarkers and support the development of radiomics- and artificial intelligence–based decision support tools for canine adrenal disease. In this context, reliable automated segmentation represents an essential prerequisite for future studies aimed at improving diagnosis, prognostic stratification, and treatment monitoring.

Despite these promising findings, several limitations of the present study should be acknowledged. First, manual segmentations served as the reference standard. Although interobserver agreement was explicitly evaluated, delineation of adrenal structures remains inherently subjective, particularly when the boundary between normal adrenal tissue and nodular lesions is poorly defined. Second, although the dataset was collected from multiple institutions and included considerable variability in

scanners and acquisition protocols, external validation on an independent cohort was not performed. Finally, interobserver variability was evaluated using two readers, with disagreements resolved through expert review during dataset preparation. Although this approach ensured a consistent reference standard, inclusion of additional independent observers could further strengthen the assessment of human variability.

Overall, this study demonstrates that deep learning–based segmentation of canine adrenal glands and adrenal nodules is feasible in a heterogeneous multicentric dataset and can achieve agreement approaching that of expert observers. These findings provide a solid foundation for the future development of automated quantitative imaging tools to support veterinary adrenal imaging in both clinical practice and research.

Methods

Ethics statement: This retrospective study included CT scans performed between June 2019 and July 2024. Ethical approval was waived due to the retrospective nature of the study, and informed consent for personal data processing was obtained from the owners. All experiments were conducted in accordance with relevant guidelines and regulations.

Dataset

Case selection: The dataset was assembled to establish a comprehensive database for the development and evaluation of a deep learning model based on the U-Net architecture for automated segmentation of adrenal glands and adrenal nodules in canine CT examinations. The collected data were intended to support the training, validation, and testing phases of the algorithm. To enhance model generalizability and reduce potential selection bias, CT examinations acquired using numerous scanners were included. CT studies were retrospectively retrieved from the databases of Antech Imaging Services , and the Ospedale Veterinario Universitario Didattico (O.V.U.D.) of the University of Padua, for the selected study period. Cases were initially identified based on radiology reports describing the presence of adrenal nodules. All examinations were subsequently reviewed using Horos (Horos Project, USA) to extract the delayed post-contrast total-body or abdominal phase and assess eligibility.

CT studies were excluded if any of the following criteria were present:

- Large abdominal masses causing marked displacement of surrounding organs.
- Moderate to severe abdominal effusion or peritoneal metastases in contact with target organs, preventing reliable anatomical delineation.
- Motion artifacts due to respiratory movement.
- Poor image quality (e.g., poor spatial resolution preventing reliable adrenal delineation).

After applying the exclusion criteria, the remaining CT examinations were included in the study and used for model development and evaluation, while a separate subset was reserved for inter-observer agreement analysis.

Ground truth determination: Manual segmentation was performed using the open-source software 3D Slicer (version 5.10.0). Segmentations were performed slice-by-slice on all images in which the target structures were visible. Two radiologists with different levels of expertise in tomographic image segmentation performed the annotations (B.S., 8 years of experience in veterinary diagnostic imaging; M.N., 1 year of experience in veterinary diagnostic imaging), which were subsequently reviewed by T.B., who has 15 years of experience in veterinary diagnostic imaging. When adrenal nodules were present, segmentation was performed according to their radiological appearance. If a clear interface between the nodular tissue and the residual adrenal parenchyma was visible, the two components were segmented separately. Adrenal masses invading adjacent major vessels were included in the dataset; however, intravascular tumor components were excluded from segmentation due to inconsistent delineation boundaries. When clearly distinguishable, the phrenicoabdominal artery was excluded from the adrenal gland segmentation.

Model architecture

The nnU-Net v2 framework[24] was employed for model development, owing to its self-configuring capabilities and demonstrated effectiveness across diverse medical image segmentation tasks. This framework automatically adapts its U-Net–based architecture[20]—including preprocessing, network topology, and training hyperparameters—to the characteristics of the input data, thereby reducing the need for manual intervention.

We employed the Residual Encoder (ResEnc) variant in the large (L) configuration[25]. This architecture replaces the plain convolutional blocks in the encoder with residual blocks, which stabilize gradient flow and facilitate training of deeper networks. The L preset is optimized for modern GPUs with approximately 22 GB VRAM and offers larger patch sizes and higher feature map dimensionality compared with the smaller M preset.

Four model configurations were trained and compared:

1. **2D model:** Processes individual axial slices independently, without inter-slice context. This configuration is computationally efficient and serves as a baseline.
2. **3D low-resolution model:** Operates on downsampled 3D volumes, capturing spatial context across multiple slices while reducing GPU memory requirements.
3. **3D full-resolution model:** Handles volumetric patches directly at the native target resolution, maintaining fine spatial detail in all three dimensions. This configuration provides the best balance between anatomical fidelity and computational cost, but comes with an increased risk of overfitting.
4. **3D cascading full-resolution model:** 3D full-resolution model run with prediction initialization from 3D low-resolution model.

To reduce the nodule false positives produced by the segmentation model, we added a lightweight binary classifier applied as a second stage. For each gland it received the source CT cropped to the adrenal-plus-nodule region delineated by the segmentation output and predicted whether that region contained a nodule, allowing predicted nodules in negative regions to be discarded. Running the classifier first could have lowered the overall computational cost, as the segmentation network is the larger of the two; however, because the adrenal glands are small and the segmentation model delineated them reliably, we used segmentation as a preprocessing step that provided the classifier's region of interest.

Training configuration

Preprocessing: Cases were resampled to a uniform voxel spacing automatically determined by the nnU-Net framework based on the dataset statistics (Tab. 6). Raw image intensities were normalized using a statistics-based procedure: the mean, standard deviation, and the 0.5th and 99.5th percentiles of the intensities within the segmentation masks were computed. Intensities were then clipped to the percentile thresholds, the mean was subtracted, and the result was divided by the standard deviation.

Table 6. Preprocessing configuration and patch sizes for each model variant. Spacings and patch sizes for the segmentation models were determined automatically by the nnU-Net v2 framework; those for the classification model were derived from the dataset median spacing and the adrenal bounding box plus a fixed margin.

Model	Spacing [mm]	Patch Size [voxels]
Classification (EfficientNet-B3)	2.0, 0.555, 0.555	$64 \times 192 \times 224$
2D	0.555, 0.555	512×512
3D Low-Resolution	2.0, 0.864, 0.864	$160 \times 192 \times 224$
3D Full-Resolution	2.0, 0.555, 0.555	$112 \times 256 \times 256$

Patch sampling: Throughout training, the network processed randomly sampled patches from the resampled volumes. This approach captures sufficient anatomical context within GPU memory constraints and simultaneously acts as data augmentation. To mitigate class imbalance—given the small volume of adrenal glands relative to the full abdominal scan—33.3% of training patches were constrained to include the foreground class, while the remaining 66.7% were sampled from random locations within the volume. This is in line with most studies utilizing the nnU-Net framework.

Loss function and optimization: For optimization, we employed a combined loss consisting of the batch DSC loss and cross-entropy loss. With batch DSC, the DSC is computed as if all samples in a mini-batch constitute a single volume, thereby stabilizing the gradient signal in cases where individual patches contain few or no lesion voxels. This is particularly relevant for adrenal segmentation, where the target structures occupy a minimal fraction of the total volume.

Optimization was performed using stochastic gradient descent (SGD) with an initial learning rate of 1×10^{-2}, a weight decay of 3×10^{-5}, and Nesterov momentum of 0.99. A polynomial learning rate decay schedule was applied throughout training.

Cross-validation: A 5-fold cross-validation approach was employed. In each fold, 80% of the 295 training cases were used for training and 20% for validation. For each of the three model configurations, five individual models were trained (one per fold). During inference, the predictions of all five fold models were combined by averaging the softmax logits, yielding a single ensemble prediction.

Minimum volume threshold: A configurable minimum volume threshold was applied during post-processing: any connected component in the predicted segmentation with a volume below 0.1 ml was discarded (replaced by background). This step was intended to suppress spurious small-volume false positives that may arise due to noise or partial-volume effects.

Secondary classification for false-positive reduction: The classifier was a 3D EfficientNet-B3 (MONAI), trained from scratch on single-channel CT patches that were resampled and intensity-normalized following the same statistics-based

procedure used for segmentation and cropped to a fixed region of interest around each gland, with the left and right adrenals treated as separate samples. It was trained with 5-fold cross-validation using a binary cross-entropy loss, the AdamW optimizer with a cosine-annealing schedule, mixed precision, and class-balanced sampling, retaining the checkpoint with the best validation AUROC, together with standard nnU-Net-style image augmentation. A SwinUNETR initialized from CT-pretrained weights[26,27] and an EfficientNet-B6 were also evaluated but performed worse and were more prone to overfitting.

Hardware and software: All training and inference were conducted on Intel Xeon E5-2698 CPU, with *251 GiB* of system memory. GPU acceleration was provided by *NVIDIA Tesla V100 DGXS 32GB* with *300 W* power draw, running CUDA version *12.8*. The software environment comprised Python *3.11.14*, PyTorch *2.9.0*, nnU-Net v2 *2.6.2*, SimpleITK *2.5.2*, and *Ubuntu 22.04.5 LTS*.

Evaluation metrics

Model performance was assessed using four complementary metrics that jointly characterize both volumetric overlap and boundary accuracy:

- **Dice Similarity Coefficient (DSC):** Measures the voxel-wise overlap between the predicted segmentation and the ground truth. DSC was computed for each individual case and then averaged across the test set. This per-case averaging weights all cases equally, making the metric sensitive to performance on small structures.
- **Aggregated Dice Score ($\mathbf{DSC_{Agg}}$):** Computed by accumulating all true positives (TP), false positives (FP), and false negatives (FN) across the entire test set before calculating DSC. This formulation naturally gives more influence to larger structures, providing a more balanced assessment that reduces the disproportionate impact of very small lesions on the overall score.
- **Average Symmetric Surface Distance (ASSD):** Measures the mean bidirectional distance between the surfaces of the predicted and ground truth segmentations. ASSD provides a stable estimate of typical boundary alignment.
- **95th-Percentile Hausdorff Distance (HD95):** The 95th percentile of the directed Hausdorff distances, capturing extreme boundary errors while reducing sensitivity to isolated outliers.

For structured analysis, metrics were calculated for two grouped label categories: (1) *organs* (left and right adrenal glands) and (2) *nodules* (left and right adrenal nodules). Metrics were also computed per label for left and right structures to evaluate performance differences by laterality.

Inter-observer agreement study

Quantifying human variability and comparative performance: Inter-observer variability was quantified by computing DSC, $\mathrm{DSC_{Agg}}$, ASSD, and HD95 between Expert A and Expert B for each case and structure group (whole adrenals and nodules). The resulting metrics define the "human envelope"—the baseline range of disagreement between trained experts. For each of the 37 cases and each model configuration, three pairwise comparisons were then computed:

- Expert A vs. Expert B
- Model vs. Expert A
- Model vs. Expert B

This three-way design determines whether the model's agreement with each individual expert falls within the range of inter-expert agreement.

Statistical analysis: For each model configuration and structure group, the distribution of pairwise DSC values between the model and each expert was compared against the expert–expert DSC distribution using the Wilcoxon signed-rank test, with statistical significance set at $p < 0.05$. A non-significant result indicates that the model–expert agreement is not discernible from inter-expert variability. The reliability of volumetric measurements among the three observers was assessed using a two-way random-effects intraclass correlation coefficient with absolute agreement (ICC(2,1)). Volumetric agreement between the model and the mean of the two expert segmentations was further evaluated using Bland–Altman analysis, reporting the mean bias and 95% limits of agreement.

Results were visualised as side-by-side boxplots comparing DSC, HD95, and ASSD distributions across the three pairwise comparison groups, and as Bland–Altman plots, for each model configuration and structure group.

References

1. Pagani, E. *et al.* Ultrasonographic features of adrenal gland lesions in dogs can aid in diagnosis. *BMC Vet. Res.* **12**, DOI: 10.1186/s12917-016-0895-1 (2016).

2. Lunn, K. F. & Boston, S. E. Tumors of the endocrine system. In Vail, D. M., Thamm, D. H. & Liptak, J. M. (eds.) *Withrow and MacEwen's Small Animal Clinical Oncology*, 565–596, DOI: 10.1016/B978-0-323-59496-7.00026-8 (Elsevier, 2020), 6 edn.

3. Cook, A. K., Spaulding, K. A. & Edwards, J. F. Clinical findings in dogs with incidental adrenal gland lesions determined by ultrasonography: 151 cases (2007–2010). *J. Am. Vet. Med. Assoc.* **244**, 1181–1185, DOI: 10.2460/javma.244.10.1181 (2014).

4. Baum, J. I., Boston, S. E. & Case, J. B. Prevalence of adrenal gland masses as incidental findings during abdominal computed tomography in dogs: 270 cases (2013–2014). *J. Am. Vet. Med. Assoc.* **249**, 1165–1169, DOI: 10.2460/javma.249.10.1165 (2016).

5. Burti, S. *et al.* Contrast-enhanced ultrasound features of adrenal lesions in dogs. *Vet. Rec.* **193**, DOI: 10.1002/vetr.2949 (2023).

6. Park, S.-H., Cho, H., Lee, K. & Yoon, H. Deep learning model for ct-based adrenal gland volume determination and normal reference definition in dogs. *Pak. Vet. J.* **45**, 320–327, DOI: 10.29261/pakvetj/2025.138 (2025).

7. Kliewer, M. A., Brinkman, M. R. & Hinshaw, J. L. The back alleys and dark corners of abdomen and pelvis computed tomography: The most frequent sites of missed findings in the multiplanar era. *J. Clin. Imaging Sci.* **10**, 70, DOI: 10.25259/jcis_184_2020 (2020).

8. Ahn, C. H. *et al.* Two-stage deep learning model for adrenal nodule detection on ct images: A retrospective study. *Radiology* **314**, DOI: 10.1148/radiol.231650 (2025).

9. Meng, F. *et al.* A deep learning algorithm for automated adrenal gland segmentation on non-contrast ct images. *BMC Med. Imaging* **25**, DOI: 10.1186/s12880-025-01682-5 (2025).

10. Kim, T. M. *et al.* Automated deep learning for detection and measurement of adrenal masses in contrast-enhanced abdominal ct. *Eur. Radiol.* DOI: 10.1007/s00330-025-12314-5 (2026).

11. Xiao, S. *et al.* Review of applications of deep learning in veterinary diagnostics and animal health. *Front. Vet. Sci.* **12**, DOI: 10.3389/fvets.2025.1511522 (2025).

12. Burti, S. *et al.* Artificial intelligence in veterinary diagnostic imaging: Perspectives and limitations. *Res. Vet. Sci.* **175**, DOI: 10.1016/j.rvsc.2024.105317 (2024).

13. Banzato, T. *et al.* Automatic classification of canine thoracic radiographs using deep learning. *Sci. Rep.* **11**, 3964 (2021).

14. Banzato, T. *et al.* An ai-based algorithm for the automatic evaluation of image quality in canine thoracic radiographs. *Sci. Reports* **13**, DOI: 10.1038/s41598-023-44089-4 (2023).

15. Dourson, A., Taylor, K., Qiao, X. & Fitzke, M. Vet-dino: Learning anatomical understanding through multi-view distillation in veterinary imaging (2025). 2505.15248.

16. Tahghighi, P., Appleby, R. B., Norena, N., Ukwatta, E. & Komeili, A. Classification of the quality of canine and feline ventrodorsal and dorsoventral thoracic radiographs through machine learning. *IEEE Access* DOI: 10.1111/vru.13373 (2024).

17. Adrien-Maxence, H. *et al.* Comparison of error rates between four pretrained densenet convolutional neural network models and 13 board-certified veterinary radiologists when evaluating 15 labels of canine thoracic radiographs. *Vet. Radiol. Ultrasound* 1–13, DOI: 10.1111/vru.13069 (2022).

18. Burti, S., Osti, V. L., Zotti, A. & Banzato, T. Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs. *The Vet. J.* **262**, 105505, DOI: 10.1016/j.tvjl.2020.105505 (2020).

19. Celniak, W. *et al.* Improving the classification of veterinary thoracic radiographs through inter-species and inter-pathology self-supervised pre-training of deep learning models. *Sci. Rep.* **13**, 19518 (2023).

20. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *arXiv preprint* (2015). 1505.04597.

21. Qureshi, I. *et al.* Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Inf. Fusion* **90**, 316–352 (2023).

22. Robinson-Weiss, C. *et al.* Machine learning for adrenal gland segmentation and classification of normal and adrenal masses at ct. *Radiology* **306**, DOI: 10.1148/radiol.220101 (2023).

23. Kim, T. M. *et al.* Fully automatic volume measurement of the adrenal gland on CT using deep learning to classify adrenal hyperplasia. *Eur. Radiol.* **33**, 4292–4302 (2023).

24. 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, DOI: 10.1038/s41592-020-01008-z (2021).

25. Isensee, F. *et al.* nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In Linguraru, M. G. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 488–498, DOI: 10.1007/978-3-031-72114-4_47 (Springer Nature Switzerland, Cham, 2024).

26. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266* (2022).

27. Tang, Y. *et al.* Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740 (2022).

Acknowledgements

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2026/019392. The research was funded in part by the Ministry of Science and Higher Education under subsidy, AGH reference number 10000-501.00-120000.

Funding

This work was supported by a research grant from the Department of Animal Medicine, Production and Health (MAPS), University of Padua, Italy (SID Burti 2024; €15,000; "Application of artificial intelligence algorithms for the study of renal parenchyma in relation to the development of chronic kidney disease in dogs and cats").

Author contributions statement

S.B. and T.B. conceived and designed the study. S.B. and N.M. collected the data, performed the manual segmentations, interpreted the results, and drafted the manuscript. A.J. developed the deep learning pipeline, trained the models, and performed the computational and statistical analyses. T.B. supervised the study, reviewed the manual segmentations, contributed to data interpretation, and critically revised the manuscript. All authors reviewed and approved the final manuscript.

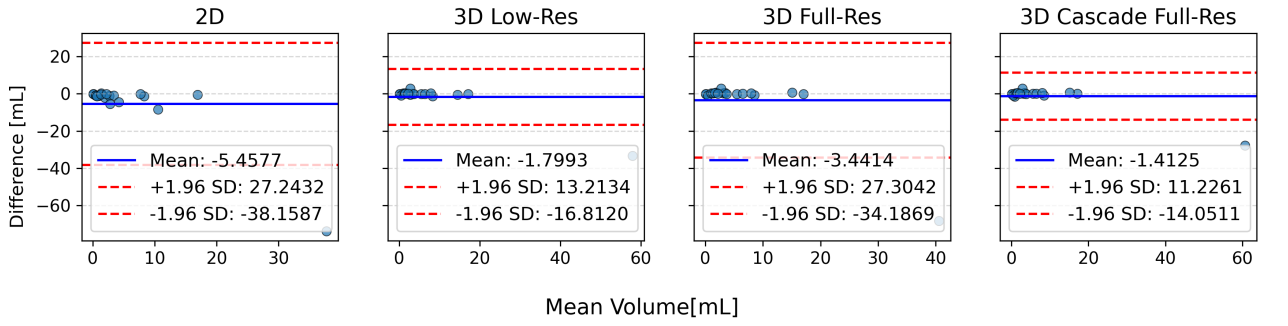
Data availability

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

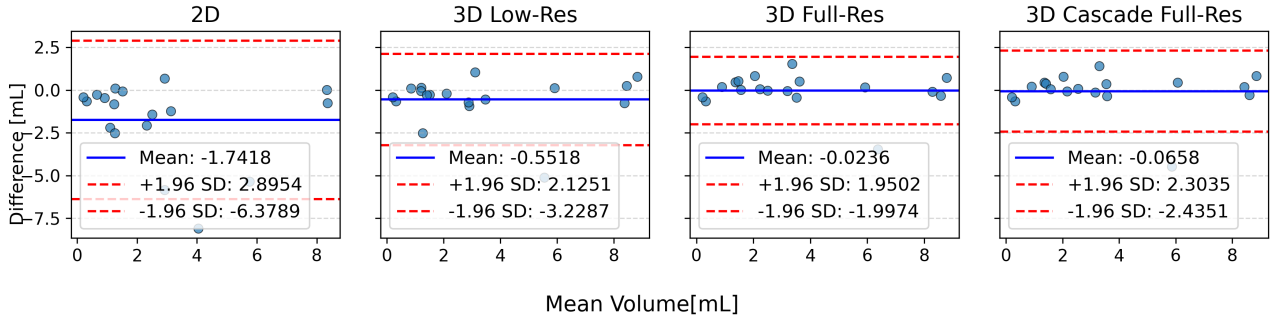
Additional information

Competing interests

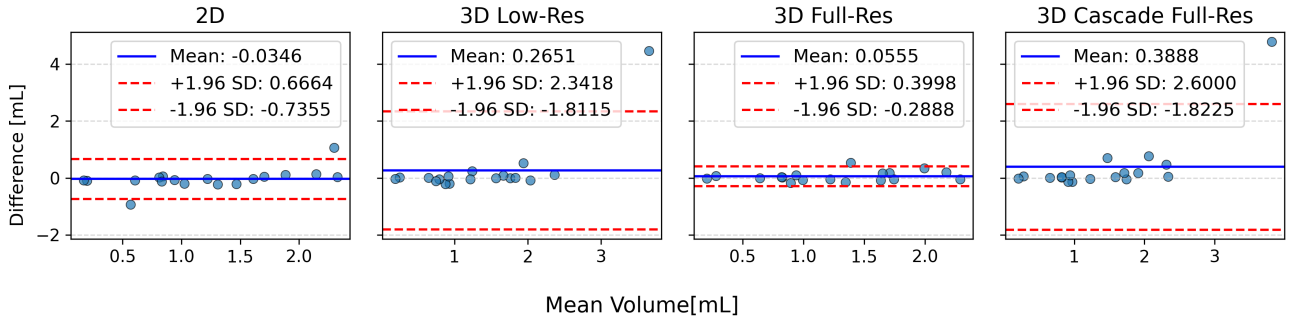
The authors declare no competing interests.



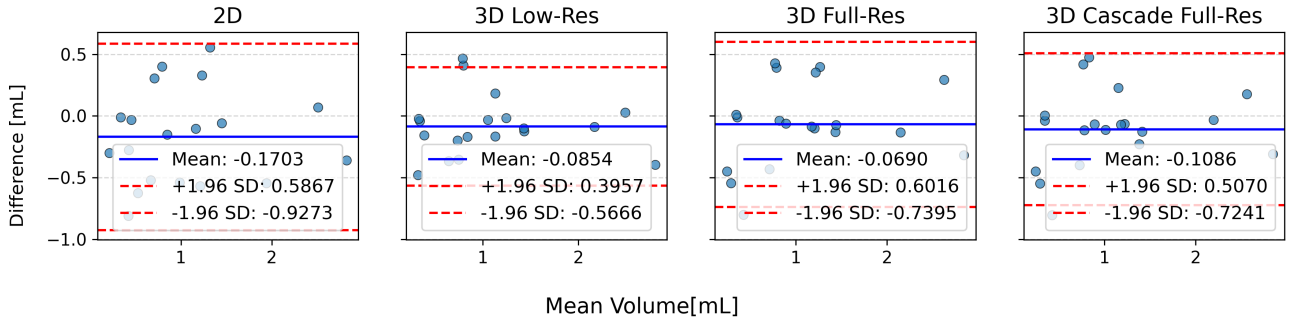
(a) Left nodules



(b) Right nodules

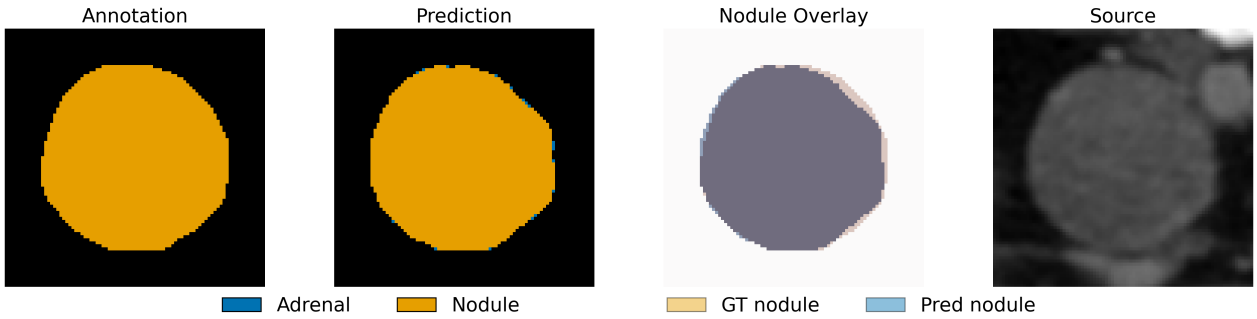


(c) Left normal adrenals

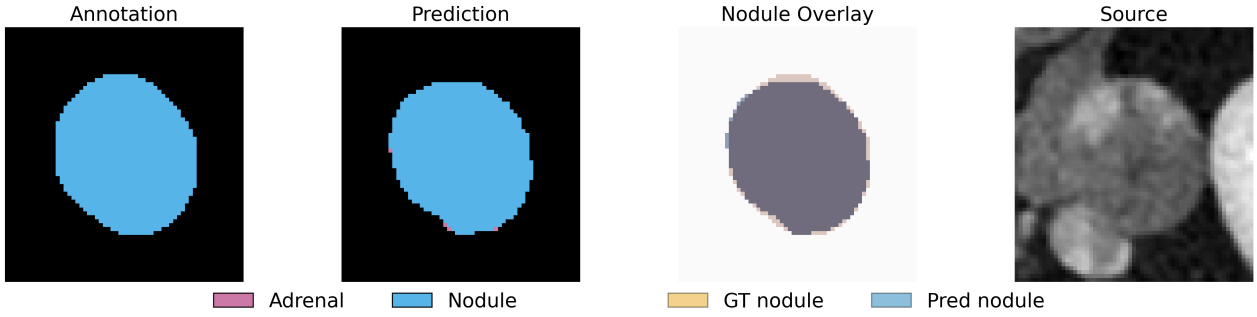


(d) Right normal adrenals

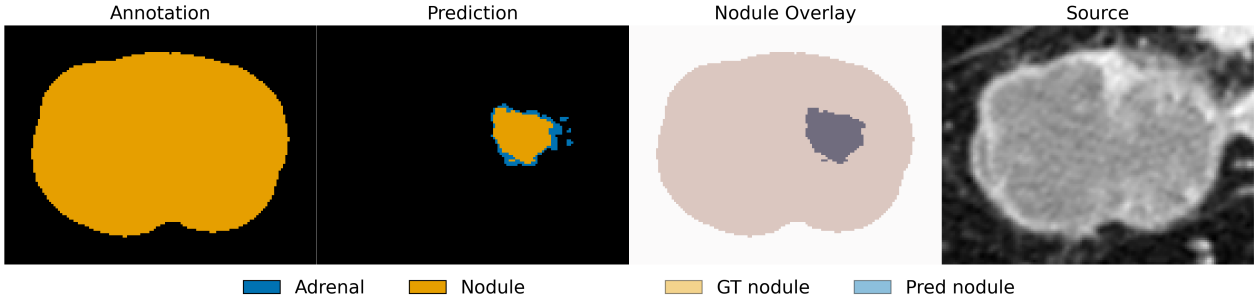
Figure 2. Bland–Altman plots comparing model-predicted volumes against the mean of the two expert segmentations, shown separately for left and right nodules and normal adrenal glands. Differences are computed as model-predicted minus mean expert volume, so negative values indicate under-segmentation and positive values over-segmentation. The solid blue line indicates the mean bias; dashed red lines indicate the 95% limits of agreement. Note, for the left normal adrenals, the positive bias and wider limits of agreement of the 3D low-resolution and 3D cascade full-resolution models, reflecting occasional over-segmentation of the gland.



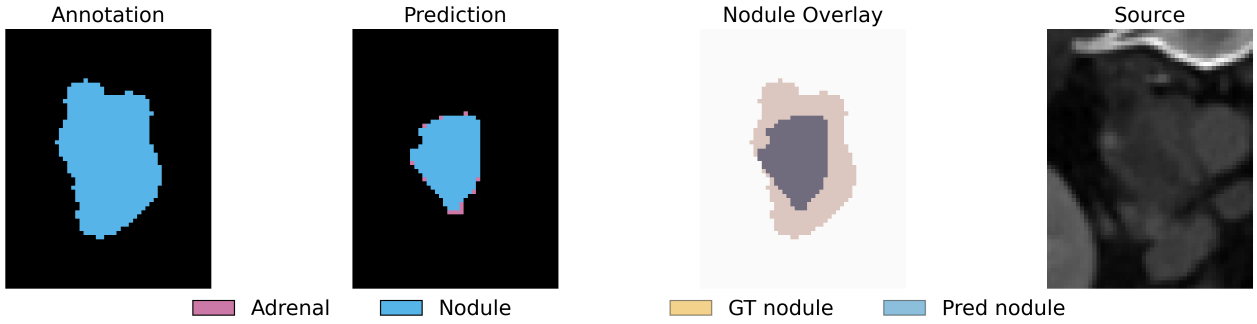
(a) Stable left adrenal gland with a nodule, demonstrating a DSC of 0.97.



(b) Stable right adrenal gland with a nodule, demonstrating a DSC of 0.93.



(c) More severely affected left adrenal gland containing a nodule. DSC is 0.13.



(d) More severely affected right adrenal gland containing a nodule. DSC is 0.66.

Figure 3. Visualization of the best- and worst-performing cases, ordered by mean DSC relative to experts A and B. The displayed slice corresponds to the one containing the largest nodule area. DSC = Dice similarity coefficient