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.

CNN model-based image classification for canine brain MRI abnormalities

Samira Abani

samira.abani@icloud.com

University of Veterinary Medicine Hannover

Merlin Laue

DOS Software-Systems

Peter J Dickinson

University of California, Davis

Steven De Decker

Royal Veterinary College

Rodrigo Gutierrez-Quintana

School of Biodiversity, One Health and Veterinary Medicine, University of Glasgow

Siavash Ghiasvand

TU Dresden

Sahar Tahmasebi

Leibniz University Hannover

Franziska Spohn

University of Veterinary Medicine Hannover

Alexander Sabbotin

University of Veterinary Medicine Hannover

Alexej Hänsch

DOS Software-Systems

Jörg Janisch

DOS Software-Systems

Reiner Ulrich

Leipzig University

Ehren McLarty

University of California, Davis

Jasmin Nessler

University of Veterinary Medicine Hannover

Holger Volk

University of Veterinary Medicine Hannover

Article

Keywords: Artificial intelligence (AI), Deep learning, Magnetic resonance imaging (MRI), Brain MRI, Abnormality detection, Diagnostic imaging

Posted Date: September 11th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7537077/v1

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

Additional Declarations: Competing interest reported. Authors J. Janisch, A. Hansch, and M. Laue are employed by DOS Software-Systeme GmbH, a commercial developer of AI solutions. The employer, DOS Software-Systeme GmbH, had no role in the interpretation of the data or in the conclusions drawn from this study. S. Abani, J. Nessler, A. Hansch, and M. Laue were funded by the “Central Innovation Program for Small- and Medium-Sized Enterprises” of the German Federal Ministry for Economic Affairs and Climate Action (grant no. KK5066602LB). Additionally, we acknowledge financial support from the Open Access Publication Fund of the University of Veterinary Medicine Hannover, Foundation. Both funders had no influence on the study design, data collection and analysis, or the conclusions drawn in this paper. The remaining authors declare no competing interests.

Abstract

Diagnostic imaging represents one of the most promising clinical applications of artificial intelligence (AI) algorithms. The purpose of this study was to evaluate the customised convolutional neural network (CNN) *SepNetDense* for distinguishing between normal and abnormal canine brain magnetic resonance images (MRIs), with the aim of enhancing diagnostic efficiency and assisting radiologists in identifying abnormal images.

The dataset comprised T1-weighted (T1w) pre- and post-contrast sequences in transverse, sagittal, and dorsal planes from 550 dogs, collected from four universities. Dogs were included if they had a complete clinical diagnosis confirmed either through histopathology or in accordance with current clinical consensus. Patients were randomly divided into a training set (n = 444), a validation set (n = 53), and a test set (n = 53). Each MRI was labelled on a slice-by-slice basis as normal or abnormal. The model was trained on 205 normal imaging datasets (e.g., extracranial aetiologies, idiopathic epilepsy, paroxysmal dyskinesia) and 239 abnormal ones (e.g., neoplasms, inflammatory lesions, other pathologies).

The model correctly predicted 74% of the true normal slices in the test set as normal and 73% of the true abnormal as abnormal. A ROC analysis of the model's prediction at the patient level revealed that, at a threshold of 51% abnormal slices per patient, the model reached an optimal balance of 83% sensitivity and 78% specificity, with a maximal accuracy of 80%. ANCOVA revealed that the CNN's classification performance was influenced by multiple biological, technical, and institutional factors. Lesion diagnosis, institutional setting, and breed size had significant large effects, whereas body weight showed a significant medium effect. Significant interactions with large effect sizes were also observed between diagnosis and institute, as well as between breed size and weight. Additionally, a distinct pattern was observed in the distribution of prediction categories along the anatomical axes, with a trend towards better CNN performance in the central quartiles across all MRI sequences. The T1w pre-contrast sagittal sequence demonstrated the highest classification accuracy (81.8%) compared to the other T1w sequences. This study evaluates a CNN-based model designed to support a triage system for classifying canine brain MRI studies, with the aim of identifying abnormalities and improving reporting efficiency.

Introduction

In recent decades, technical innovations in veterinary diagnostic imaging, especially magnetic resonance imaging (MRI), have profoundly advanced clinical practice in veterinary neurology and neurosurgery [1]. MRI serves as a non-invasive, highly sensitive tool for detecting abnormalities, tracking disease progression, and evaluating treatment and post-surgical outcomes [2,3]. Serial MRI studies further enable in vivo monitoring of lesion progression in real time, making MRI an essential and versatile tool for the antemortem assessment of canine brain pathologies [3]. The demand for imaging services, not only in human but also in veterinary medicine, has outpaced the number of trained diagnostic imaging specialists, creating a widening gap between the availability of imaging technology and diagnostic expertise [1,4].

While the number of MRI units in universities and private practices continues to rise, the workforce of veterinary radiologists and neurologists has not increased at the same rate, posing challenges for maintaining high-quality image interpretation [1,5]. In recent years, convolutional neural networks (CNNs) have become a powerful tool in medical imaging for tasks such as classification, detection, localisation, segmentation, augmentation, and automated diagnosis [6-9].

CNNs are particularly effective, as they can automatically learn complex patterns and hierarchical features from image data by adjusting millions of parameters through optimisation [10,11]. The architecture of a CNN typically comprises three main components: the input layer, the hidden layers, and the output layer. The hidden layers include convolutional layers, which use filters to extract key features from images; pooling layers, which reduce dimensionality; and fully connected layers, which make final predictions. This ability to process and learn from visual data makes CNNs particularly valuable for medical imaging applications [11]. Numerous studies have demonstrated that CNNs can assist clinicians in a range of image analysis applications across different anatomical regions, achieving performance levels comparable to those of human experts [12-19]. While CNNs continue to advance in human medicine, improving accuracy, robustness, and generalisability, their adoption in veterinary diagnostic practice remains at an early stage [20,21].

In 2018, several studies began to explore the potential of CNNs for analysing veterinary image diagnostic [22-28]. Among veterinary diagnostic imaging modalities, radiography is the most widely studied due to its common use in routine veterinary practice [20,29]. MRI, particularly in the context of neurological conditions, is an emerging area of increasing research interest but remains comparatively underexplored relative to radiography [20,29].

Niemeyer et al. (2024) fine-tuned a deep learning tool (VGG-16 network), originally developed for the automatic grading of human lumbar intervertebral disc (IVD) degeneration using the Pfirrmann scheme, to apply the same grading system to T2-weighted (T2w) midsagittal images of lumbar canine spines. The model achieved an average accuracy of 94.1%, a sensitivity of 85.2%, and a specificity of 96.3% for grades 1 to 5, demonstrating the potential to advance both veterinary care and human biomedical research [30].

In another study, Biercher et al. (2021) demonstrated the potential of CNNs in differentiating thoracolumbar spinal cord diseases in dogs. The CNN model achieved a sensitivity of 90.80% and a specificity of 98.98% for intervertebral disc extrusion (IVDE) using T2w sagittal images; a sensitivity of 100% and a specificity of 95.10% for intervertebral disc prolapse (IVDP) using sagittal T1-weighted (T1w) images; and a sensitivity of 90.98% and a specificity of 90.12% for fibrocartilaginous embolism/acute non-compressive nucleus pulposus extrusion (FCE/ANNPE) using T2w transverse images. This study emphasises the potential of CNNs to provide second opinions in the assessment of spinal cord lesions on MRI, while also highlighting the challenges posed by limited training data for certain diagnoses [31].

Since 2018, the use of CNNs for analysing canine brain tumours using MRI has gained increasing attention in veterinary neuroimaging [25]. Banzato et al. (2018) utilised two neural networks (NNs), one

based on AlexNet and the other a customised model, to predict the histopathological grading of canine meningiomas. Their preliminary study tested the models on ten randomly selected images, correctly classifying the meningioma grade in eight out of ten cases. This result underscores the potential of MRI-based tumour grading in guiding treatment decisions [25]. Later that year, Banzato et al. trained a CNN based on GoogLeNet to distinguish between meningiomas and gliomas on canine MRI images [23]. The model performed best on post-contrast T1w images, achieving an accuracy of 94%. Performance was slightly lower on pre-contrast T1w and T2w images, with respective accuracies of 91% and 90%. These findings further demonstrate the potential of deep learning models to assist in the diagnosis and classification of various tumour types [23].

While recent studies have highlighted the growing use of CNNs in veterinary imaging diagnostics, their application in triaging or enhancing diagnostic workflows for complex imaging modalities, such as brain MRI, remains limited. This is partly due to the lack of heterogeneous large datasets that adequately reflect real-world variability. The hypothesis of this study is that a CNN model trained on a heterogeneous brain MRI dataset can accurately predict whether a dog's brain is normal or abnormal. To test this hypothesis, we: (1) annotated 79,310 MRI slices from 550 dogs across four institutions, comprising both unremarkable brains as well as brains with multiple diseases; (2) trained the customised CNN model, *SepNetDense*, using the training and validation sets; and (3) evaluated the model's performance on an independent test set.

Material and Methods

Dataset and Features. The dataset comprised T1w sequences acquired pre- and post-contrast in the transverse, sagittal, and dorsal planes. It included 241 patients with unremarkable brain imaging findings (e.g., extracranial aetiologies, idiopathic epilepsy, and paroxysmal dyskinesia), and 309 patients with remarkable imaging findings (e.g., neoplasms, inflammatory lesions, and other pathological abnormalities). Dogs were included if they had a complete clinical diagnosis, confirmed either by histopathology or in accordance with current clinical scientific consensus, which served as the ground truth.

Breed, age, weight, sex, presenting complaint, imaging findings, and final diagnoses were extracted from the electronic medical records of each patient. All MRI studies were reviewed by at least one veterinary neuroradiology expert. No animals were directly involved in this study, as all data were retrospectively collected from clinical cases that had previously undergone brain MRI for diagnostic purposes.

Patient Selection. Electronic databases from two university referral centres, the Department of Small Animal Medicine and Surgery at the University of Veterinary Medicine Hannover, Germany (TiHo), and the William R. Pritchard Veterinary Medical Teaching Hospital at the University of California, Davis (UC Davis), were searched for dogs presenting with clinical signs of encephalopathy between April 2003 and June 2022. A total of 178 patients with MRI studies showing remarkable brain findings were included if

they had a comprehensive clinical history and received a histopathological diagnosis from a board-certified pathologist, based on either post-mortem examination or stereotactic brain biopsy.

In addition, 84 cases with histopathologically confirmed primary or secondary intracranial neoplasia, from the dataset of a previous study conducted at the Royal Veterinary College, Small Animal Referral Hospital, London (RVC), were included [32]. Furthermore, 45 cases of meningoencephalitis of unknown origin, with a final or presumptive diagnosis as reported in a previous study conducted at the Small Animal Hospital of the University of Glasgow (Glasgow), were incorporated [33].

MRI studies were acquired at four university referral centres using clinical MRI scanners from Philips, GE Medical Systems, and Siemens. At TiHo, imaging was performed using 3.0 T scanners, including the Philips Achieva and Achieva dStream models. UC Davis used 1.5 T scanners from GE Medical Systems, including the Genesis Signa and Signa HDxt. Glasgow acquired scans using 1.5 T systems, including the Philips NT Intera and Siemens Magnetom Essenza. At RVC, studies were obtained using various 1.5 T scanners, including the Philips Achieva, Gyroscan, Intera, NT Intera, and the GE Genesis Signa. Different imaging protocols were applied at each institution. Differences in scanner models, magnet strength, and acquisition protocols across institutions were considered as part of the dataset's heterogeneity.

Control Group. The control group consisted of MR images from 241 dogs with unremarkable brain findings, obtained from four university referral centres. This group included dogs with conditions such as extracranial aetiologies; paroxysmal dyskinesia, diagnosed in accordance with the consensus statement of the European College of Veterinary Neurology (ECVN) [34]; and idiopathic epilepsy, classified as tier II according to the International Veterinary Epilepsy Task Force consensus statement [35].

As individual dogs may have undergone multiple MRI examinations on different dates at the referral centres, only one MRI study per case was included in the dataset. Follow-up MRI studies performed post-surgery or post-treatment were excluded. Approval for the use of medical data was obtained through the owner's informed consent prior to hospital admission, in accordance with each university's institutional guidelines. The acquisition of MRI images was supervised by veterinary technicians from the diagnostic imaging units of the referral centres, in compliance with their respective animal welfare policies. All patients were imaged under general anaesthesia. MRI studies affected by severe motion artefacts or incomplete sequences that prevented diagnostic interpretation were excluded from the study dataset.

Dataset Preparation. All MR images were anonymised using Python™ (Python™, version 3.11. Python Software Foundation, DE, USA) by removing metadata tags containing personal information [36].

Labelling. All images were in Digital Imaging and Communications in Medicine (DICOM) format and analysed using available DICOM viewers: RadiAnt™ (RadiAnt™ DICOM Viewer for Windows, version 4.0.1. Poznań, Poland) or OsiriX MD (OsiriX MD, DICOM viewer for macOS, Pixmeo SARL, Geneva, Switzerland) [37, 38]. The DICOM files were extracted and subsequently converted into PNG format. MRI slices were labelled slice by slice using the online annotation platform V7 Darwin (V7 Labs. V7 Darwin, London, UK) [39].

Brain MR images were annotated using the following tags, which served as the ground truth: **(a) normal**, indicating the absence of detectable abnormalities and serving as a baseline for comparison; and **(b) pathological structural changes**, including lesions, brain oedema, mass effect, ventricular abnormalities, midline shift, parenchymal loss, pathological contrast enhancement, and herniation. For binary classification, images labelled as *normal* were categorised as normal, while those tagged with any pathological condition were classified as abnormal.

Experimental Design and Data Split. To train, validate, and test the performance of the presented CNN, the dataset comprising 550 dogs was randomly divided into a training set, a validation set, and a test set. Randomisation was performed using MATLAB® (MATLAB® software, MathWorks Inc. version R2022b. Natick, MA, USA), without accounting for covariates such as age, weight, breed type, institution, number of sequences, or disease category [40]. For data splitting, all slices and sequences from each patient were treated as a single unit, and the dataset was split at the patient level. The model was trained at the slice level, where normal brain MRI images consisted exclusively of normal slices, while remarkable cases contained both abnormal and normal slices, as the lesion was not present in every slice.

Training set: This set consisted of 444 dogs and included 39,208 normal and 11,960 abnormal images, enabling the model to learn and adjust its parameters based on these data. **Validation set**: The validation set comprised 53 dogs, with 5,491 normal and 1,860 abnormal slices. It was used to fine-tune the model's hyperparameters and to monitor performance during the training process. **Test set**: The test set also consisted of 53 dogs and included 4,612 normal and 1,989 abnormal imaging slices. This set was reserved for final model evaluation, providing an unbiased estimate of the model's performance and ensuring its ability to generalise to unseen data.

Pre-processing. The initial T1w image file sizes ranged from 179 bytes to 399.91 KB, with matrix dimensions varying between 144 × 144 and 1024 × 1024 pixels. Pixel dimensions were dependent on the scanner's resolution and field of view at the time of acquisition. All T1w images (greyscale) were resized to 128 × 128 pixels, and normalised during pre-processing. To increase data variability, augmentation techniques were applied, including random 90° rotations and horizontal or vertical flips.

CNN model. Although well-established networks such as ResNets and DenseNets have demonstrated strong performance across a range of vision tasks, they performed suboptimal for the present application [41, 42]. Due to their high parameter count, these models tended to either overfit or underfit the data, failing to capture domain-specific features essential for robust classification [41, 42]. To address this limitation, a customised, optimised architecture, *SepNetDense*, was developed for binary brain classification in the present study. *SepNetDense* features a small parameter footprint, comprising a total of 76,626 parameters, of which 65,746 are trainable [43–45]. The architecture of *SepNetDense* is illustrated in Fig. 1.

The following classification metrics were calculated:

$$Classification\ Accuracy\ =\ (TP\ +\ TN)\ /\ (TP\ +\ TN\ +\ FP\ +\ FN)$$

1

$$Precision = TP / (TP + FP)$$

2

$$Recall = TP / (TP + FN)$$

3

$$F1 - score = 2 \times Precision \times Recall / (Precision + Recall)$$

4

TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

Frequency distributions of the predictions were compared using chi-square tests or Fisher's exact tests and visualized as heatmaps using GraphPad Prism (Prism 8 for Windows, version 8.4.3, GraphPad Software, San Diego, CA, USA) [46].

The proportion of slices predicted to be normal or abnormal as compared to all slices at the patient level was assessed for normality using a Q-Q plot and Shapiro-Wilk tests, and analysed for differences between the groups using the Mann–Whitney U test. Receiver Operating Characteristic (ROC) analysis was performed to assess classification performance and determine optimal thresholds, also using GraphPad Prism (Prism 8 for Windows, version 8.4.3, GraphPad Software, San Diego, CA, USA).

An analysis of covariance (ANCOVA) was performed in SPSS (IBM SPSS Statistics for Windows, version 29.0.1.1, IBM Corp., Armonk, NY, USA) to examine and compare the effects of materials- and methods-related factors, covariates, and selected interactions on classification accuracy [47]. Levene's test was applied to assess the assumption of homogeneity of variances. Effect sizes were reported using partial eta squared (η^2), interpreted as small (0.01–0.059), medium (0.06–0.13), and large (≥ 0.14), as described by Gravetter et al. [48].

A two-way analysis of variance (ANOVA) with post hoc one-way ANOVAs and subsequent pair-wise post hoc tests with Bonferroni corrections was applied to assess the effects of the categorical factors diagnosis and institute on the CNN model's accuracy using SPSS (IBM SPSS Statistics for Windows, version 29.0.1.1, IBM Corp., Armonk, NY, USA). Levene's test was applied to assess the assumption of homogeneity of variances and normality was assessed using a Q-Q plot of standardized residues [49]. The influence of the categorical factors institute, manufacturer, scanner model, field strength, diagnosis, breed size, skull conformation and gender on accuracy was analysed individually. Distributions were first assessed with Levene's test, then groups were compared using one-way ANOVA or Kruskal-Wallis tests with Bonferroni post hoc corrections, or unpaired t-tests or Mann-Whitney U tests where applicable. Linear regression analyses were used to assess associations between the continuous factors and classification accuracy using GraphPad Prism (Prism 8 for Windows, version 8.4.3, GraphPad Software,

San Diego, CA, USA). To evaluate the adequacy of the linear regression models, residual plots were examined.

In general, statistical significance was defined as a $p \leq 0.05$.

Results

Dataset. The study included the MRI sequences of 550 canine brains with pre- and post-contrast T1w images, collected across four referral centres. From these patients, a total of 79,310 MRI slices were annotated manually as either normal or abnormal, based on brain MRI findings. Slices affected by artefacts or considered to be of insufficient quality were excluded, resulting in a final dataset comprising 65,120 slices. The training set consisted of 444 patients, of which 205 had normal and 239 had abnormal brain findings. The validation and test sets each comprised 53 patients, with 18 normal and 35 abnormal patients.

In the study dataset (Table 1), the most represented diagnosis was the MRI-normal group (n = 241), comprising patients with unremarkable brain MRI findings. This was followed by the neoplastic group (n = 186) and the inflammatory group (n = 113). The least represented was the other causes group (n = 10), which included vascular, hereditary, and degenerative brain abnormalities.

Table 1
Distribution of diagnostic categories across the training, validation, and test sets (n = 550 patients).

Diagnostic Category	Training set	Validation set	Test set
MRI-Normal	205 (46.17%)	18 (33.96%)	18 (33.96%)
Neoplastic	143 (32.21%)	21 (39.62%)	22 (41.51%)
Inflammatory	89 (20.05%)	12 (22.64%)	12 (22.64%)
Other causes	7 (1.58%)	2 (3.77%)	1 (1.89%)
Total	444	53	53

The study dataset represented a heterogeneous mix of breeds (**Suppl. Table 1**). The most represented breeds were crossbreeds (16%), followed by Labrador Retrievers (11%) and Boxers (5%). The mean age of the study group was 5.81 ± 3.60 years, and the mean body weight was 22.55 ± 14.06 kg. MRI studies included in the dataset were acquired using 1.5 T (n = 309) and 3.0 T (n = 241) scanners. The most commonly used manufacturers were Philips (n = 363), GE Medical Systems (n = 123), and Siemens (n = 64).

Performance. Using the described test set (n = 53; Table 1), the performance of the CNN was evaluated in classifying individual MRI slices as either normal or abnormal. A total of 6,660 MRI slices were included in the analysis. Based on manual annotations serving as the reference standard, 73%

(1,452/1,989) of abnormal slices and 74% (3,430/4,612) of normal slices were correctly classified, whereas 27% (537/1,989) of abnormal slices and 26% (1,182/4,612) of normal slices were incorrectly classified (Fig. 2).

Based on individual MRI slice classification, the model achieved a precision of 0.86, recall of 0.74, and F1-score of 0.80 for normal slices, and a precision of 0.55, recall of 0.73, and

F1-score of 0.63 for abnormal slices, with an overall accuracy of 0.74 (Table 2).

Table 2
Performance metrics

Class	Precision	Recall	F1-score
Normal	0.86	0.74	0.80
Abnormal	0.55	0.73	0.63
Overall accuracy			0.74

Findings. The spatial distribution of classification errors (false positives and false negatives) was analysed slice-wise using heatmaps to visualize prediction patterns across all orientations in pre- and post-contrast T1w sequences after normalisation (Fig. 3). Across all three orientations; transverse (Fig. 3a–b), sagittal (Fig. 3c–d), and dorsal (Fig. 3e–f) T1w images of the normal patient subset are characterized by a high number of true negatives across slices. False positives appear as scattered patches along the slice axes, predominantly in the central portions, and vary between individual patients. In the abnormal subset, true positives dominate the central portions of the slice axes across all orientations, forming dense horizontal bands, while false negatives occur more sporadically in the same regions. By contrast, false positives and true negatives tend to cluster near the periphery of the anatomical axes, particularly at the sequence extremes. In dorsal pre- and post-contrast T1w images (Fig. 3e–f), true positive clusters in abnormal patients are generally shorter and less continuous than in other orientations. Error patterns in this view appear more fragmented, which might suggest localized deviations and greater uncertainty in the dorsal plane. To investigate whether the prediction outcomes (true positive, true negative, false positive, false negative) varied systematically along the anatomical axes, each MRI sequence was normalised slice-wise and divided into four equal quartiles (Q1–Q4). Results showed that the distribution of the prediction categories differed significantly across the quartiles in all sequences (**Suppl. Table 2**).

Slice-wise accuracy across T1w sequences. The relationship between CNN slice-wise accuracy and MRI sequence type was examined. As illustrated in Fig. 4, a significant difference was observed in the accuracy achieved with the different sequences (Chi-square test, $p \leq 0.001$). Pre-contrast T1w sagittal images demonstrated the highest slice-wise accuracy (81.8%), significantly outperforming other sequences, including pre-contrast T1w transverse (71.2%), post-contrast T1w transverse (73.3%), post-

contrast T1w sagittal (75.1%), post-contrast T1w dorsal (76.1%), and pre-contrast T1w dorsal sequence (69.7%).

Assessment of the sensitivity and specificity of the CNN model at the patient level using receiver operating characteristic analysis (ROC). Since the CNN's predictions were based on individual MRI slices, the ROC curve was used to evaluate the model's performance at the patient level, determining the optimal threshold for classifying patients as abnormal or normal, based on the proportion of slices predicted to be abnormal slices among all slices.

The proportion of slices predicted to be abnormal in the normal (n = 18) and abnormal (n = 35) patient groups in the test set was assessed using a Q-Q plot (**Suppl. Figure 1**), and Shapiro–Wilk tests, which indicated significant deviations from normality in both groups (normal $p = 0.027$; abnormal $p = 0.003$).

Figure 5a shows that the proportion of slices predicted to be abnormal within each patient differs between the groups. The median of the proportion of slices predicted to be abnormal was significantly higher in the abnormal group at 0.70, compared to 0.26 in the normal group (Mann–Whitney U test $p < 0.0001$).

The ROC curve is presented in Fig. 5b. The area under the curve (AUC) was 0.85 ($p < 0.0001$), indicating that the model reliably distinguishes between normal and abnormal patients.

An optimal balance between sensitivity and specificity was achieved when a threshold of 51% predicted abnormal slices out of all slices was used to classify a patient as abnormal. This threshold resulted in a sensitivity of 83%, specificity of 78%, and a positive likelihood ratio of 3.7, corresponding to the highest overall accuracy of 80%.

A more conservative threshold of 64% maximised specificity to 94%, while maintaining a reasonable sensitivity of 66%. This resulted in the highest positive likelihood ratio of 11.8, with the model's accuracy at this threshold being 80%. Detailed performance metrics at each threshold are presented in **Suppl. Table 3.**

Identification of Parameters Influencing CNN Model Performance Using ANCOVA. In the next step, a univariate ANCOVA was conducted to evaluate the influence of materials- and methods-related parameters on the accuracy of the CNN model. The included categorical factors were institute (TiHo, UC Davis, RVC, Glasgow), breed size (small, medium, large), skull conformation (brachycephalic vs non-brachycephalic), sex (male, female), diagnosis (MRI-normal, neoplastic, inflammatory, other causes), and the quantitative covariates were age, weight, number of slices, and the number of available MRI sequences per case in the test set (**Suppl. Table 4**). In addition, interactions between diagnosis and institute, breed size and weight, as well as sex and weight were examined. As each university referral centre employed distinct imaging protocols, which differed in MRI settings, equipment, and acquisition techniques, the factor ‘institute’ was treated as a composite factor representing the combined influence of these imaging-related variations on model accuracy. Scanner model and magnetic field strength were

not included in the ANCOVA due to multicollinearity with the variable “institute,” as each institute in our dataset used a unique scanner. Attempts to substitute “manufacturer” with “scanner model” did not resolve this issue, since each model was still exclusive to one institute. Therefore, “institute” was retained in the model to account for all scanner- and site-specific factors, allowing for interpretable results without redundancy.

The corrected model was statistically significant ($p < 0.001$, $\eta^2 = 0.80$). Levene’s test rejected the assumption of variance homogeneity ($p = 0.021$), likely due to the large number of factor levels relative to the total sample size (n = 53). Since the Q–Q plot showed only minor deviations from normality, the model was retained, and its results were interpreted with caution **(Suppl. Figure 2)**. Effect sizes were calculated using partial eta squared (η^2) and interpreted as small, medium, or large, as mentioned in the Materials and Methods section (Fig. 6). As shown in Fig. 6a, the ANCOVA indicated that several parameters significantly influenced the accuracy of the CNN model. Main effects with large effect sizes were observed for breed size ($p < 0.001$, $\eta^2 = 0.44$), institute ($p = 0.001$, $\eta^2 = 0.39$), and diagnosis type ($p = 0.003$, $\eta^2 = 0.35$). Medium effect sizes were identified for body weight ($p = 0.050$, $\eta^2 = 0.12$). Significant interaction effects with large effect sizes were found between diagnosis and institute ($p < 0.001$, $\eta^2 = 0.45$), as well as between breed size and weight ($p < 0.001$, $\eta^2 = 0.43$). In contrast, sex, sex and weight interaction, skull conformation, age, number of sequences per case, and number of slices per case exhibited no significant effect on the CNN-models accuracy ($p > 0.05$, η^2 ranging from 0.065 to 0.004).

Following the identification of the main factors influencing model performance through ANCOVA, a two-way ANOVA was conducted to further investigate the effects of institute and diagnosis on model accuracy (Fig. 6b). The analysis revealed significant main effects of institute ($p < 0.001$) and diagnosis ($p = 0.007$), as well as a significant interaction ($p < 0.001$), indicating that the effect of diagnosis on model performance was dependent on the institute. A pair wise post hoc comparison of the institutes reveals higher accuracy with TiHo data as compared to RVC and Glasgow ($p < 0.001$ and $p = 0.001$), as well as higher accuracy with UC Davis data as compared to Glasgow ($p = 0.029$). The two-way ANOVA has to be interpreted carefully, since Levene’s test rejected the assumption of variance homogeneity ($p = 0.014$). However, visual inspection of the Q-Q plot of the standardized residuals shows approximate normality (**Suppl. Figure 3**), and Levene’s test for all post hoc one-way ANOVAs did not reject the assumption of variance homogeneity (TiHo $p = 0.199$; UC Davis $p = 0.148$; RVC $p = 0.813$; Glasgow $p = 0.075$). Accordingly, one-way ANOVA revealed a significant effect of diagnosis in the TiHo subset (ANOVA, $p = 0.028$), with MRI-normal cases achieving significantly higher accuracy than inflammatory cases (Bonferroni correction, $p = 0.031$). Furthermore, a significant effect of diagnosis was found in the RVC cohort (ANOVA, $p = 0.040$); however, pairwise post hoc tests could not be calculated due to the small number of cases in the ‘other causes’ group. In contrast, no significant effects of diagnosis on accuracy were detected in the UC Davis (ANOVA, $p = 0.055$) and Glasgow subsets (ANOVA, $p = 0.143$).

Due to the significant interaction between weight and breed size identified in the ANCOVA, the effect of body weight on CNN classification accuracy was examined separately within each breed size category

(small, medium, large) using linear regressions (Fig. 6c; residual plot in **Suppl Fig. 4)**. In small breeds, no significant association was found between weight and accuracy ($p = 0.6006$, $R^2 = 0.0165$). Similarly, the association was not significant in medium breeds ($p = 0.3390$, $R^2 = 0.0916$). However, in large breeds, a significant negative association was observed ($p = 0.0186$, $R^2 = 0.2471$), indicating that higher body weight within the large breeds was associated with reduced CNN accuracy.

In addition to the ANCOVA, the individual effects of categorical factors on CNN accuracy were examined using one-factor tests (**Suppl. Figure 5**). Institution significantly affected accuracy, with TiHo cases showing higher accuracy than RVC and Glasgow (Levene's test, $p = 0.016$; Kruskal-Wallis test, $p \leq 0.001$; **Suppl. Figure 5a**). Accuracy was not significantly affected by manufacturer (Levene's test, $p = 0.022$; Kruskal-Wallis test, $p = 0.119$; **Suppl. Figure 5c**). Scanner model had a significant effect, with Achieva showing higher accuracy than Intera and Magnetom Essenza (Levene's test, $p = 0.050$; Kruskal-Wallis test, $p = 0.003$; **Suppl. Figure 5e**). Scanner field strength significantly influenced accuracy, with 3 T datasets showing higher accuracy than 1.5 T datasets (Levene's test, $p = 0.052$; independent t-test, $p < 0.001$; **Suppl. Figure 5g**). Accuracy was not significantly affected by breed size (Levene's test, $p = 0.131$; ANOVA, $p = 0.400$; **Suppl. Figure 5i**), skull conformation (Levene's test, $p = 0.288$; independent t-test, $p = 0.374$; **Suppl. Figure 5k**), sex (Levene's test, $p = 0.877$; independent t-test, $p = 0.815$; **Suppl. Figure 5m**), or diagnosis (Levene's test, $p = 0.001$; Kruskal-Wallis test, $p = 0.848$; **Suppl. Figure 5o**).

Simple linear regression analyses were performed to assess whether CNN classification accuracy was individually influenced by age at MRI, weight at MRI, number of available sequences, and number of slices per case (**Suppl. Figure 6**). The number of sequences demonstrated a positive linear relationship with accuracy ($p = 0.003$, $R^2 = 0.162$; **Suppl. Figure 6a**). Similarly, the number of slices was positively linearly correlated with accuracy ($p < 0.001$, $R^2 = 0.222$; **Suppl. Figure 6b)**. In contrast, no significant linear correlation was found for weight at MRI ($p = 0.839$, $R^2 = 0.001$; **Suppl. Figure 6e**) and age at MRI ($p = 0.853$, $R^2 = 0.001$; **Suppl. Figure 6g**).

Discussion

CNN models are emerging as promising diagnostic tools in medical imaging, though their application in veterinary medicine remains in its early stages. In this study, we evaluated a CNN model to classify canine brain pre- and post-contrast T1w MRI studies as normal or abnormal, aiming to enable computer-assisted detection of brain lesions and improve reporting efficiency in clinical practice. Beyond evaluating overall performance, we examined the influence of biological, technical, and institutional factors on model accuracy. Our ANCOVA analysis revealed that lesion diagnosis, institutional setting, and breed size each had significant effects, all with large effect sizes, while body weight showed a medium effect. Significant interactions with large effect sizes were also observed between diagnosis and institute, as well as between breed size and weight. The model achieved its highest accuracy on the T1w pre-contrast sagittal sequence. Moreover, a spatial pattern was observed along the three anatomical axes, with higher performance in central brain regions and more classification errors in peripheral slices. These findings underscore the multifactorial nature of CNN performance and highlight

the importance of dataset diversity, anatomical context, and sequence selection in developing robust and generalisable AI models for veterinary neurology.

The model was trained on 444 MRI data sets collected from four veterinary referral centres, including a variety of breeds, ages, body weights, diagnoses, and MRI scanner types. This multi-institutional dataset was designed to reflect real-world variability and to support the development of a model capable of generalising across diverse clinical settings. Previous studies exploring CNNs in veterinary imaging have provided valuable insights; however, they were limited by small cohorts or single-centre dataset [22–28]. In contrast, our study utilises a broader and more heterogeneous dataset that better represents the complexity of clinical practice.

The customised CNN model presented in this study demonstrated a good ability to classify canine brain MRI slices as normal or abnormal. At the slice level, it achieved an overall accuracy of 74%, with comparable classification rates for normal (74%) and abnormal (73%) slices (Fig. 2). As shown in Table 2, the model achieved a precision of 0.86 for normal slices, indicating that predictions of normality were generally reliable. In contrast, the lower precision for abnormal slices of 0.55 reflects a higher rate of false positives in this class. Nonetheless, recall values were similar for both classes (0.74 for normal and 0.73 for abnormal), indicating that the model successfully identified the majority of true instances. The corresponding F1-scores were 0.80 for normal and 0.63 for abnormal slices. This discrepancy may be attributed to the relatively limited dataset size compared with the complexity of the classification task, which may have constrained the model's exposure to the full range of abnormal presentations [50]. Furthermore, the heterogeneous and often subtle nature of abnormalities likely contributed to the increased false positive rate [6, 50].

To date, few studies have used CNN models to evaluate canine brain MRI. Banzato et al. reported high classification accuracies in CNN-based canine brain MRI studies, achieving up to 94% when distinguishing between meningiomas and gliomas [23], and correctly predicting meningioma grade in 80% of cases [25]. The discrepancy in accuracy between our study and previous applications of CNNs in veterinary neurology is likely attributable to differences in sample size and methodological design. Importantly, small sample sizes in neuroimaging studies employing machine learning have been shown to introduce substantial variability in performance estimates, thereby limiting the reliability of conclusions and the generalisability of findings [51]. These studies focused on well-defined tumour types and employed smaller and simpler datasets, which may have simplified the classification task. In contrast, our model was trained on a larger, multi-institutional dataset that included a broader spectrum of diagnoses and greater variability in breed, MRI scanner type, and acquisition protocols. Hence, based on the ANCOVA results, the intentional inclusion of real-world variability in our experimental design is likely the primary factor contributing to the lower accuracy observed compared to the previous studies. Our analysis shows that the almost unlimited variety of possible diagnoses and the variability of the institutional framework conditions are the biggest factors influencing the performance of the CNN model. While this heterogeneity may result in the numerically lower performance as compared to the previously mentioned studies, the inclusion of real-life variability is a prerequisite for the development of

practical applications, since it offers a more realistic representation of clinical diversity. Training on low-variability data sourced from a single institution carries the risk of poor reproducibility, as models may end up capturing dataset-specific artefacts rather than learning generalisable features [52]. Zhang et al. (2016) showed that deep neural networks are capable of fitting random labels or noise, emphasising their potential to memorise data without extracting meaningful patterns [52]. Consequently, models that demonstrate strong performance under narrowly controlled research settings often fail to translate to real-world scenarios, highlighting the need for large, heterogeneous datasets and thorough external validation to ensure robust generalisation. Ultimately, while this heterogeneity may result in numerically lower performance compared to the previous studies, incorporating real-life variability is essential for the development of practical applications, as it provides a more accurate reflection of clinical diversity [23, 25].

The ANCOVA indicated that CNN classification performance was influenced by several biological, technical, and institutional factors, which may need to be considered as confounding factors in future studies (Fig. 6). Imaging-related differences, including MRI protocols, manufacturer, scanner model, magnetic field strength, and acquisition techniques, summarised as 'institute,' along with the range of morphological changes represented by 'diagnosis', are shown in our results to be the major factors affecting CNN model performance. The CNN model performed best on cases from TiHo, and poorest on cases from Glasgow, reflecting the different number of cases from the institutes in the training dataset. Lower accuracy for a particular institute or diagnostic category indicates suboptimal performance on that subset, but does not imply that these cases are inherently more challenging for the CNN model. This finding aligns with previous studies that underscore the susceptibility of CNN-based models to domain shifts in medical imaging [53, 54]. Variations in patient demographics, imaging protocols, and institutional practices present significant challenges for cross-site generalisation, ultimately affecting the model’s ability to perform consistently across different clinical environments [53, 54].

A critical comparison of the ANCOVA results with the one-factor analyses shows that, although multiple factors individually influence accuracy, their effects are negligible compared to the main factors, which are institutional setting, diagnostic category, breed size, and weight, when the full experimental design is considered. Since CNN-based models learn patterns from the data they are exposed to, it is crucial to train them on diverse datasets from multiple institutes and a variety of diseases, ensuring the model can distinguish between different classes in the test set [55, 56]. Recent research has shown that differences in patient populations, imaging methods, and institutional protocols can lead to substantial domain shifts, which consequently hinder the generalisability of CNN-based models across diverse clinical settings [53, 54]. To address this, training on diverse, multi-centre datasets that incorporate a wide range of pathologies is essential to achieve reliable model performance in independent, real-world environments [52, 55].

In this study, a slice-by-slice approach was employed for CNN training, resulting in predictions made on individual MRI slices. To interpret these predictions at the patient level, we explored the optimal threshold for classifying a patient as abnormal or normal based on the proportion of slices predicted to

be abnormal versus all slices within the respective patient. The results indicated that there is no single optimal threshold; rather, two thresholds (a balanced threshold of 51% and a more conservative 64%) may be applicable depending on the clinical context. For triage applications, where the model serves as a secondary reader or “second pair of eyes,” a more sensitive threshold of 51% may be preferable to minimise the risk of missing potential abnormalities. In contrast, for situations where high specificity is critical, such as to avoid unnecessary referrals or follow-ups, the 64% threshold may be more appropriate. Thus, the classification threshold should be adjusted according to the specific clinical objective.

As shown in Fig. 3, a pattern is observed in the distribution of prediction categories (true positive, true negative, false positive, false negative) along the anatomical axes of the slices within all sequences. This pattern is confirmed by statistical analysis, which revealed that the CNN model's performance varied significantly across quartiles (Q1–Q4) within each MRI sequence. The true positive category is most prominent in the central region of the slice axes for abnormal patients, indicating that the model’s predictions are more stable in these mid-slice regions. One possible explanation is the greater anatomical homogeneity of central brain structures across patients, allowing the model to learn more generalisable and discriminative features. Furthermore, central slices contain a larger volume of brain tissue, providing richer spatial context, which may enhance model confidence and prediction accuracy.

In contrast, a higher density of classification errors is observed near the extremes of the slice axes. This may be due to reduced structural information, increased anatomical variability, inconsistent signal characteristics, or a combination of biological and imaging-related factors. Peripheral slices typically contain less brain tissue and a greater proportion of non-neural structures, such as bone, cerebrospinal fluid, or air-filled spaces. These regions are more prone to partial volume effects and imaging artefacts, both of which can degrade feature quality and reduce the model’s predictive reliability.

Based on these findings, we concluded that the relative position of each slice along the anatomical axes influences prediction consistency, with this effect varying by sequence and orientation. To address these limitations, Kamnitsas et al. (2017) suggested excluding peripheral slices from training due to their limited diagnostic value. In their multi-scale 3D CNN for brain lesion segmentation, they excluded outer slices, citing poor information quality at the volume boundaries [57]. Furthermore, the implementation of a robust skull-stripping step, as described by Nour Eddin et al. (2023), prior to training could reduce the influence of non-brain structures [58]. However, excluding peripheral slices should be approached with caution, as these slices are routinely reviewed in clinical practice. Incorporating peripheral slices into model training might enhance robustness and better reflect the diagnostic process.

It is worth noting that in our study, body weight was negatively associated with classification accuracy in large breeds, a pattern not observed in small or medium breeds. This is in contrast to findings in human neonatal MRI studies, where smaller brain volume in infants have been associated with reduced image quality and diagnostic performance, largely due to lower tissue contrast, increased partial volume effects, and technical limitations [59].

While CNNs do not interpret images in the same way as human experts, they identify and weight statistical patterns in the input data that are associated with specific outputs (e.g., normal vs abnormal), without possessing any semantic understanding of anatomical structures such as a “lesion” or “brain” [60]. As a result, misclassifications in peripheral slices or in large-breed dogs may arise from non-neural structures, signal artefacts, or anatomical variability that resemble features the model has learned to associate with abnormal cases. Similar findings have been reported in human imaging studies, where CNNs were shown to rely on spurious correlations rather than pathology-specific features [61]. Explainability tools such as Gradient-weighted Class Activation Mapping (Grad-CAM), which provide visual explanations by using the gradient of a target class flowing into the final convolutional layer to generate a coarse localisation map, may help determine whether the CNN is attending to biologically relevant brain regions or being misled by non-neural features such as bone, cerebrospinal fluid, or imaging artefacts in peripheral slices [62].

Simple regression analyses showed that both the number of slices and available sequences per case positively affected accuracy, indicating that more imaging data leads to better classification performance. The model benefits from access to more data per patient, capturing additional spatial context and subtle abnormalities [63].

Another finding of this study was that the T1w pre-contrast sagittal sequence demonstrated significantly higher classification accuracy (81.8%) compared to other T1w sequences. This may be because the sagittal plane provides a clearer view of midline structures, particularly the ventricular system, which can exhibit subtle morphological changes associated with underlying pathology. In human medicine, similar observations were reported by Grigas et al. (2025) in a study distinguishing individual with mild cognitive impairment from cognitively normal controls. They found that sagittal T1w slices provided the highest diagnostic performance, followed by axial and coronal planes [59]. This was likely due to the presence of informative features in midline structures such as the corpus callosum, thalamus, and lateral ventricles, which are known to undergo structural and metabolic alterations in the early stages of cognitive decline [64].

Considering all aspects, the mentioned factors can be considered potential confounders and are important to acknowledge when designing, interpreting, or applying CNN models in a clinical context.

Limitations. This study has several limitations, and key challenges remain to be addressed in future work. A primary limitation is the relatively small number of cases included. The reason for this was that we aimed to include only cases which had a very high diagnostic accuracy, e.g. all brain neoplasms included were histopathologically confirmed. This limited the dataset significantly, albeit using the dataset from large academic institutions. Although the training set was heterogeneous and sourced from multiple institutions, the total of 444 cases may be insufficient for training a complex deep learning model, potentially limiting the model’s ability to generalise across the full spectrum of pathological presentations observed in clinical practice [50]. Nonetheless, the performance achieved can be

considered acceptable given the size and diversity of the dataset, providing a valuable proof of concept for future research.

In veterinary clinical practice, it is uncommon for completely healthy dogs without clinical signs to undergo brain MRI, making the acquisition of true normal controls challenging. As a result, the control group in this study included patients with Tier II idiopathic epilepsy, paroxysmal dyskinesia, and extracranial causes, rather than dogs with entirely normal neurological status. Although these dogs showed no structural brain abnormalities on MRI, the presence of underlying neurological disease could introduce subtle changes not detectable by expert visual assessment, potentially influencing model training and performance. Future studies would benefit from a broader and more representative control population to further enhance model reliability and generalisability.

Another key limitation of this study is the exclusive use of T1w pre- and post-contrast images for model development. While T1w imaging is effective for detecting mass lesions and abnormalities associated with blood-brain barrier disruption, it is less sensitive to other important pathological changes. Conditions such as oedema, demyelination, mild inflammatory processes, early ischemic changes, and subtle structural abnormalities are often better visualised on T2w, FLAIR, or diffusion-weighted sequences [65]. In fact, some abnormalities, particularly those without significant contrast enhancement, may have been underrepresented or missed during model training. The next phase of this study will focus on incorporating additional MRI sequences to enable more comprehensive detection of a wider range of brain pathologies.

While a direct comparison with veterinary radiologists was beyond the scope of this study, it will be essential for future validation and clinical integration of such models in real-world scenarios. Evaluating the model's performance relative to expert clinicians is a critical next step to determine whether it can match or complement human expertise. These comparisons are vital for defining the model's potential role in clinical practice, whether as a decision support tool, a triage system, or an aid for less experienced clinicians.

In this study, slices with noise, motion artefacts, and incomplete imaging were removed during dataset preparation to ensure clean inputs for model training. However, in clinical practice, MRI studies submitted for automated analysis often contain motion artefacts, suboptimal image quality, or missing sequences. As a result, models developed under controlled research conditions may perform less reliably when applied to real-world clinical data. Future work should focus on enhancing model robustness to such variability by incorporating artefact detection modules, implementing quality control procedures, and training on more heterogeneous, imperfect datasets that better reflect clinical conditions.

Conclusion

Despite the limitations mentioned, the present CNN model provides a foundation for further exploration of CNN-based methods in canine brain MRI classification. This study also highlights the value of using

real-world, multi-institutional data with a broad range of diagnoses to train CNN-based models for clinical settings. Moreover, distinct patterns in classification performance along the anatomical axes highlight the need to use explainability tools to verify whether the CNN is focusing on biologically relevant brain regions. In the medical domain, such tools are essential to build trust, ensure clinical relevance, and identify potential biases in model predictions. The findings underscore the potential of deep learning to support diagnostic workflows in veterinary neuroimaging. Future work should prioritise collaborative data sharing, the use of larger and more balanced datasets, and rigorous external validation to enhance model generalisability and facilitate clinical integration.

Declarations

Competing interests

Authors J. Janisch, A. Hansch, and M. Laue are employed by DOS Software-Systeme GmbH, a commercial developer of AI solutions. The employer, DOS Software-Systeme GmbH, had no role in the interpretation of the data or in the conclusions drawn from this study. S. Abani, J. Nessler, A. Hansch, and M. Laue were funded by the “Central Innovation Program for Small- and Medium-Sized Enterprises” of the German Federal Ministry for Economic Affairs and Climate Action (grant no. KK5066602LB). Additionally, we acknowledge financial support from the Open Access Publication Fund of the University of Veterinary Medicine Hannover, Foundation. Both funders had no influence on the study design, data collection and analysis, or the conclusions drawn in this paper. The remaining authors declare no competing interests.

Funding

This work was funded by the Central Innovation Program for small- and medium-sized enterprises of the German Federal Ministry for Economic Affairs and Climate Action - grant number KK5066602LB. Open access publishing was supported by the Open Access Publication Fund of the University of Veterinary Medicine Hannover, Foundation.

Author Contribution

H.V. supervised this study. H.V., S.A. and J.N. designed the experiments. P.J.D., S.D., R.G.Q., and E.M. contributed to data curation by providing raw MRI data and patient information. S.A. and A.S. annotated the cases. S.A. and F.S. extracted the patient data. S.A., S.T., S.GH., and R.U. analysed the data. S.A. wrote the first draft of the manuscript. M.L., A.H. and J.J. developed the CNN model. All authors reviewed the manuscript and approved the final version.

Acknowledgement

The authors acknowledge the radiologists and pathologists at the University of Veterinary Medicine Hannover, Hannover, Germany; the Royal Veterinary College, London, United Kingdom; the School of Veterinary Medicine, University of California, Davis, California, USA; and the Small Animal Hospital, School of Biodiversity, One Health and Veterinary Medicine, University of Glasgow, Glasgow, United Kingdom, for providing access to and permission to use the radiology and necropsy reports for this study. We also gratefully acknowledge L. Böhringer and L. Lemke for assistance with patient selection, G. Lester, P. Hallur, D. Sanchez-Masian, and A. Wang Leonardo for their valuable consultation during the course of this study.

Data Availability

The datasets generated during the current study is available from the corresponding author upon reasonable request.

No animals were directly involved in this study, as all data were retrospectively collected from clinical cases that had previously undergone brain (MRI for diagnostic purposes. All procedures were performed as part of routine clinical care at four university veterinary referral centers: the Department of Small Animal Medicine and Surgery at the University of Veterinary Medicine Hannover (Germany), the William R. Pritchard Veterinary Medical Teaching Hospital at the University of California, Davis (USA), the Small Animal Referral Hospital at the Royal Veterinary College in London (UK), and the Small Animal Hospital of the University of Glasgow (UK). Approval for the use of medical data was obtained through the owner's informed consent prior to hospital admission, in accordance with each university’s institutional guidelines. The acquisition of MRI data was supervised by veterinary technicians from the diagnostic imaging units of the referral centres, in compliance with their respective animal welfare policies. All patients were imaged under general anaesthesia.

References

1. Gavin, P. R. Growth of clinical veterinary magnetic resonance imaging. *Vet. Radiol. Ultrasound*. **52**, S2–S4. https://doi.org/10.1111/j.1740-8261.2010.01779.x (2011).
2. Vickram, A. S., Infant, S. S., Priyanka & Chopra, H. AI-powered techniques in anatomical imaging: impacts on veterinary diagnostics and surgery. *Ann. Anat.* **258**, 152355. https://doi.org/10.1016/j.aanat.2024.152355 (2025).
3. Werring, D. J. et al. The pathogenesis of lesions and normal-appearing white matter changes in multiple sclerosis: a serial diffusion MRI study. *Brain* **123**, 1667–1676. https://doi.org/10.1093/brain/123.8.1667 (2000).
4. The Royal College of Radiologists. Clinical radiology workforce census report 2023. (2024). Available at https://www.rcr.ac.uk/media/4imb5jge/_rcr-2024-clinical-radiology-workforce-census-

report.pdf

5. Dick White Referrals. Addressing the shortage of veterinary imaging specialists. (2020). Available at https://www.dickwhitereferrals.com/news/addressing-the-shortage-of-veterinary-imaging-specialists/
6. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118. https://doi.org/10.1038/nature21056 (2017).
7. Avendi, M., Kheradvar, A. & Jafarkhani, H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med. Image Anal.* **30**, 108–119. https://doi.org/10.1016/j.media.2016.01.005 (2016).
8. Garcea, F., Serra, A., Lamberti, F. & Morra, L. Data augmentation for medical imaging: a systematic literature review. *Comput. Biol. Med.* **152**, 106391. https://doi.org/10.1016/j.compbiomed.2022.106391 (2023).
9. Bohmrah, M. K. & Kaur, H. Advanced hybridization and optimization of DNNs for medical imaging: a survey on disease detection techniques. *Artif. Intell. Rev.* **58**, 122. https://doi.org/10.1007/s10462-024-11049-x (2025).
10. Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M. & Ilono, P. Deep convolutional neural networks in medical image analysis: a review. *Information* **16**, 195. https://doi.org/10.3390/info16030195 (2025).
11. Yu, H., Yang, L. T., Zhang, Q., Armstrong, D. & Deen, M. J. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* **444**, 92–110. https://doi.org/10.1016/j.neucom.2020.04.157 (2021).
12. Liang, G. & Zheng, L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Comput. Methods Programs Biomed.* **187**, 104964. https://doi.org/10.1016/j.cmpb.2019.06.023 (2020).
13. Nahiduzzaman, M., Islam, M. R. & Hassan, R. ChestX-ray6: prediction of multiple diseases including COVID-19 from chest X-ray images using convolutional neural network. *Expert Syst. Appl.* **211**, 118576. https://doi.org/10.1016/j.eswa.2022.118576 (2023).
14. Li, L., Xu, M., Wang, X., Jiang, L. & Liu, H. Attention based glaucoma detection: a large-scale database and CNN model. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 10563–10572; (2019). https://doi.org/10.1109/CVPR.2019.01082
15. Eltoukhy, M. M., Hosny, K. M. & Kassem, M. A. Classification of multiclass histopathological breast images using residual deep learning. *Comput. Intell. Neurosci.* 9086060; (2022). https://doi.org/10.1155/2022/9086060 (2022).
16. Khened, M., Kollerathu, V. A. & Krishnamurthi, G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Med. Image Anal.* **51**, 21–45. https://doi.org/10.1016/j.media.2018.10.004 (2019).
17. Al-masni, M. A., Kim, D. H. & Kim, T. S. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Comput. Methods Programs Biomed.*

190, 105351. https://doi.org/10.1016/j.cmpb.2020.105351 (2020).

18. Lee, S. M. & Kim, N. Deep learning model ensemble for the accuracy of classification degenerative arthritis. *Comput. Mater. Continua*. **75**, 1981–1994. https://doi.org/10.32604/cmc.2023.035245 (2023).
19. Sarwinda, D., Paradisa, R. H., Bustamam, A. & Anggia, P. Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer. *Procedia Comput. Sci.* **179**, 423–431. https://doi.org/10.1016/j.procs.2021.01.025 (2021).
20. Xiao, S. et al. Review of applications of deep learning in veterinary diagnostics and animal health. *Front. Vet. Sci.* **12**, 1511522. https://doi.org/10.3389/fvets.2025.1511522 (2025).
21. Ezanno, P. et al. Research perspectives on animal health in the era of artificial intelligence. *Vet. Res.* **52**, 40. https://doi.org/10.1186/s13567-021-00902-4 (2021).
22. Banzato, T. et al. Use of transfer learning to detect diffuse degenerative hepatic diseases from ultrasound images in dogs: a methodological study. *Vet. J.* **233**, 35–40. https://doi.org/10.1016/j.tvjl.2017.12.026 (2018).
23. Banzato, T., Bernardini, M., Cherubini, G. & Zotti, A. A methodological approach for deep learning to distinguish between meningiomas and gliomas on canine MR images. *BMC Vet. Res.* **14**, 317. https://doi.org/10.1186/s12917-018-1638-2 (2018).
24. Yoon, Y., Hwang, T. & Lee, H. Prediction of radiographic abnormalities by the use of bag-of-features and convolutional neural networks. *Vet. J.* **237**, 43–48. https://doi.org/10.1016/j.tvjl.2018.05.009 (2018).
25. Banzato, T., Cherubini, G. B., Atzori, M. & Zotti, A. Development of a deep convolutional neural network to predict grading of canine meningiomas from magnetic resonance images. *Vet. J.* **235**, 90–92. https://doi.org/10.1016/j.tvjl.2018.04.001 (2018).
26. Dumortier, L., Guépin, F., Delignette-Muller, M. L., Boulocher, C. & Grenier, T. Deep learning in veterinary medicine, an approach based on CNN to detect pulmonary abnormalities from lateral thoracic radiographs in cats. *Sci. Rep.* **12**, 11418. https://doi.org/10.1038/s41598-022-14993-2 (2022).
27. Vinicki, K., Ferrari, P., Belic, M. & Turk, R. Using convolutional neural networks for determining reticulocyte percentage in cats. *arXiv preprint arXiv*. https://doi.org/10.48550/arXiv.1803.04873 (2018). :1803.04873.
28. Kim, H. J., Lee, S. H., Kim, H. J. & Lee, S. H. CNN-based diagnosis models for canine ulcerative keratitis. *Sci. Rep.* **9**, 1–10. https://doi.org/10.1038/s41598-019-50437-0 (2019).
29. Pereira, A. et al. Artificial intelligence in veterinary imaging: an overview. *Vet. Sci.* **10**, 320. https://doi.org/10.3390/vetsci10050320 (2023).
30. Niemeyer, F. et al. Automatic grading of intervertebral disc degeneration in lumbar dog spines. *JOR Spine*. **7**, e1326. https://doi.org/10.1002/jsp2.1326 (2024).
31. Biercher, A. et al. Using deep learning to detect spinal cord diseases on thoracolumbar magnetic resonance images of dogs. *Front. Vet. Sci.* **8**, 8. https://doi.org/10.3389/fvets.2021.721167 (2021).

32. Kaczmarska, A. et al. Postencephalitic epilepsy in dogs with meningoencephalitis of unknown origin: clinical features, risk factors, and long-term outcome. *J. Vet. Intern. Med.* **34**, 808–820. https://doi.org/10.1111/jvim.15687 (2020).
33. Schwartz, M., Lamb, C. R., Brodbelt, D. C. & Volk, H. A. Canine intracranial neoplasia: clinical risk factors for development of epileptic seizures. *J. Small Anim. Pract.* **52**, 632–637. https://doi.org/10.1111/j.1748-5827.2011.01131.x (2011).
34. Cerda-Gonzalez, S. et al. International veterinary canine dyskinesia task force ECVN consensus statement: terminology and classification. *J. Vet. Intern. Med.* **35**, 1218–1230. https://doi.org/10.1111/jvim.16108 (2021).
35. De Risio, L. et al. International veterinary epilepsy task force consensus proposal: diagnostic approach to epilepsy in dogs. *BMC Vet. Res.* **11**, 148. https://doi.org/10.1186/s12917-015-0462-1 (2015).
36. Python Software Foundation. Python, version 3.11. *Wilmington, DE, USA*; Available at https://www.python.org/downloads/release/python-3110/
37. Medixant *RadiAnt DICOM Viewer*, version 4.0.1. *Poznań, Poland*; Available at https://www.radiantviewer.com
38. Pixmeo, S. A. R. L. *OsiriX MD*. FDA-cleared and CE-labeled DICOM viewer for macOS. *Geneva, Switzerland*; Available at https://www.osirix-viewer.com/osirix/osirix-md/
39. V7 Labs. *V7 Darwin: AI Data Labeling Platform*. *London, UK*; (2025). Available at https://www.v7labs.com
40. The MathWorks Inc. *Statistics and Machine Learning Toolbox*, version R *Natick, MA, USA*; (2022b). Available at https://www.mathworks.com/products/statistics.html
41. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 770–778; (2016). https://doi.org/10.1109/CVPR.2016.90
42. Iandola, F. N. et al. Densenet: implementing efficient ConvNet descriptor pyramids. *arXiv preprint arXiv* :1404.1869; (2014). https://arxiv.org/abs/1404.1869
43. Chollet, F. & Xception Deep learning with depthwise separable convolutions. In Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 1800–1807. IEEE Computer Society. (2017). https://doi.org/10.1109/CVPR.2017.195
44. Howard, A. G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv* :170404861. (2017).
45. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2818–2826 (Las Vegas, NV, USA, 27–30 June 2016). (2016). https://doi.org/10.1109/CVPR.2016.308
46. GraphPad, S. *GraphPad Prism*, Prism 8 for Windows, version 8.4.3. *San Diego, CA, USA*; Available at https://www.graphpad.com

47. IBM Corp. *IBM SPSS Statistics for Windows*, version 29.0.1.1. *Armonk, NY, USA*; Available at https://www.ibm.com/products/spss-statistics
48. Gravetter, F. J. & Wallnau, L. B. *Statistics for the Behavioral Sciences* 10th edn (Cengage Learning, 2016).
49. Hoffmann, G., Lichtinghagen, R. & Wosniok, W. Ein einfaches Verfahren zur Schätzung von Referenzintervallen aus routinemäßig erhobenen Labordaten. *Lab. Med.* **39** https://doi.org/10.1515/labmed-2015-0082 (2015).
50. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011 (2018).
51. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 180, 68–77 (2018). https://doi.org/10.1016/j.neuroimage.2017.06.061. New advances in encoding and decoding of brain signals.
52. Zhang, C., Bengio, S., Hardt, M. & Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv* :161103530 (2017).
53. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLOS Med.* **15**, e1002683. https://doi.org/10.1371/journal.pmed.1002683 (2018).
54. Mahmutoglu, M. et al. Optimizing MRI sequence classification performance: insights from domain shift analysis. *Eur. Radiol.* https://doi.org/10.1007/s00330-025-11671-5 (2025).
55. Garrucho, L. et al. Domain generalization in deep learning-based mass detection in mammography: a large-scale multi-center study. *Artif. Intell. Med.* **132**, 102386. https://doi.org/10.1016/j.artmed.2022.102386 (2022).
56. Holzschuh, J. et al. The impact of multicentric datasets for the automated tumor delineation in primary prostate cancer using convolutional neural networks on 18F-PSMA-1007 PET. *Radiat. Oncol.* **19**, 106. https://doi.org/10.1186/s13014-024-02491-w (2024).
57. Kamnitsas, K. et al. DeepMedic for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 138–149 (Springer International Publishing, 2016). https://doi.org/10.1007/978-3-319-55524-9_14
58. Nour Eddin, J., Dorez, H. & Curcio, V. Automatic brain extraction and brain tissues segmentation on multi-contrast animal MRI. *Sci. Rep.* **13**, 6416. https://doi.org/10.1038/s41598-023-33289-7 (2023).
59. Dubois, J. et al. MRI of the neonatal brain: a review of methodological challenges and neuroscientific advances. *J. Magn. Reson. Imaging*. **53**, 1318–1343. https://doi.org/10.1002/jmri.27192 (2021).
60. Huff, D., Weisman, A. & Jeraj, R. Interpretation and visualization techniques for deep learning models in medical imaging. *Phys. Med. Biol.* **66** https://doi.org/10.1088/1361-6560/abcd17 (2021).
61. DeGrave, A. J., Janizek, J. D. & Lee, S. I. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv* (2020). https://doi.org/10.1101/2020.09.13.20193565 Preprint.

62. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359. https://doi.org/10.1007/s11263-019-01228-7 (2019).
63. Coupet, M. et al. A multi-sequences MRI deep framework study applied to glioma classification. *Multimed Tools Appl.* **81**, 13563–13591. https://doi.org/10.1007/s11042-022-12316-1 (2022).
64. Grigas, O., Damaševičius, R. & Maskeliūnas, R. Multimodal convolutional mixer for mild cognitive impairment detection. *Comput. Mater. Contin*. **84**, 1805–1838. https://doi.org/10.32604/cmc.2025.064354 (2025).
65. Mai, W. *Diagnostic MRI in Dogs and Cats* 1st edn (CRC, 2018).

Figures

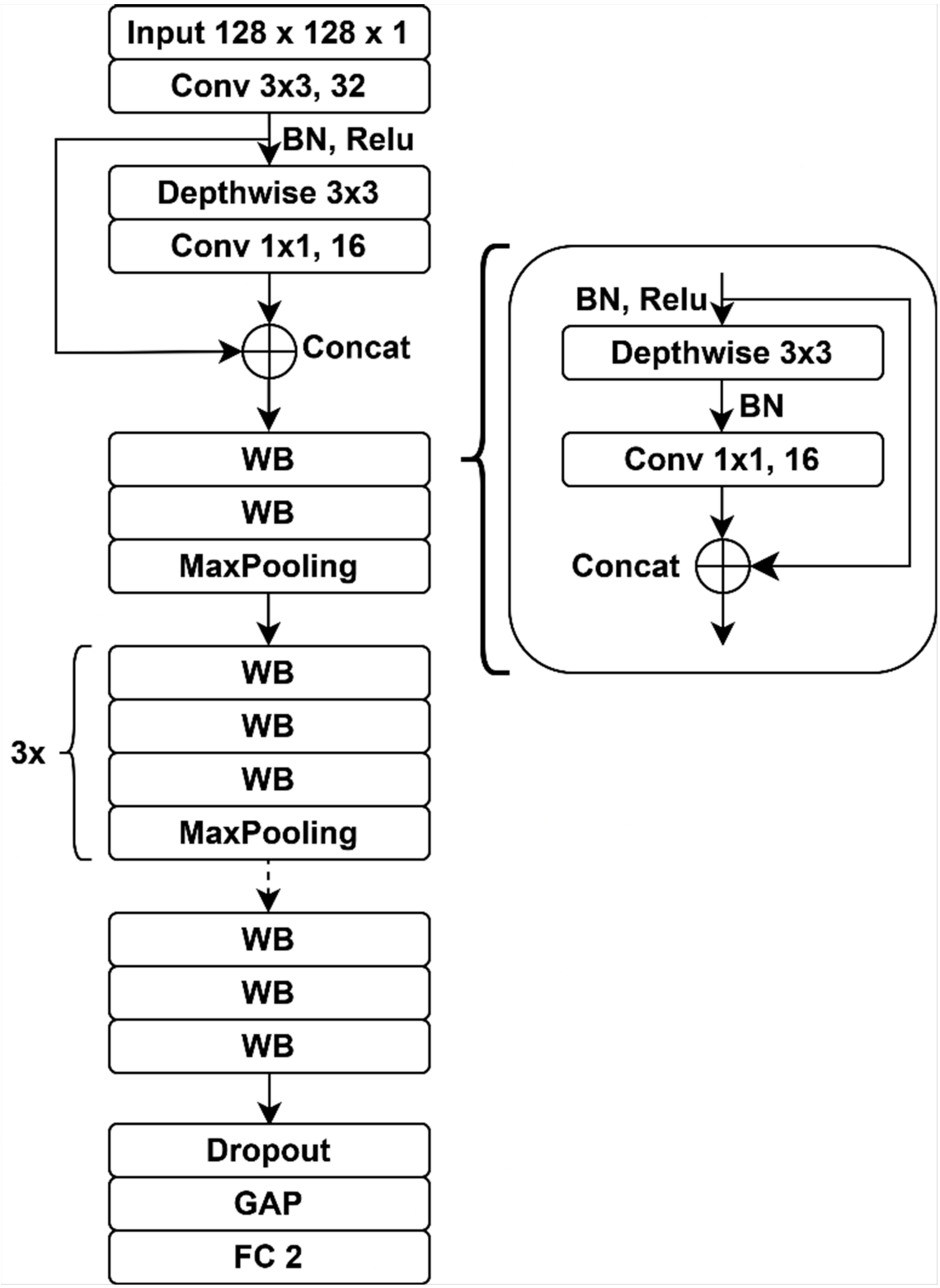


Figure 1

Schematic illustration of the *SepNetDense* model architecture. The input is a grayscale image (128×128×1) passed through a 3×3 convolutional layer (32 filters). The main building blocks of the proposed architecture are the Weight Blocks (WB), consisting of a series of depthwise separable convolution (DSC) layers, with intermediate batch normalization (BN) layers and a residual connection. The architecture of the WBs is shown on the right side of the image. Every three WBs are followed by a

MaxPooling layer. At the end, a Dropout layer prevents overfitting, followed by Global Average Pooling (GAP) and a Fully Connected (FC) layer with two output units to provide the final binary classification results. It is important to note that the intermediate BN layer from the first DSC block and the MaxPooling layer from the last WB are removed.

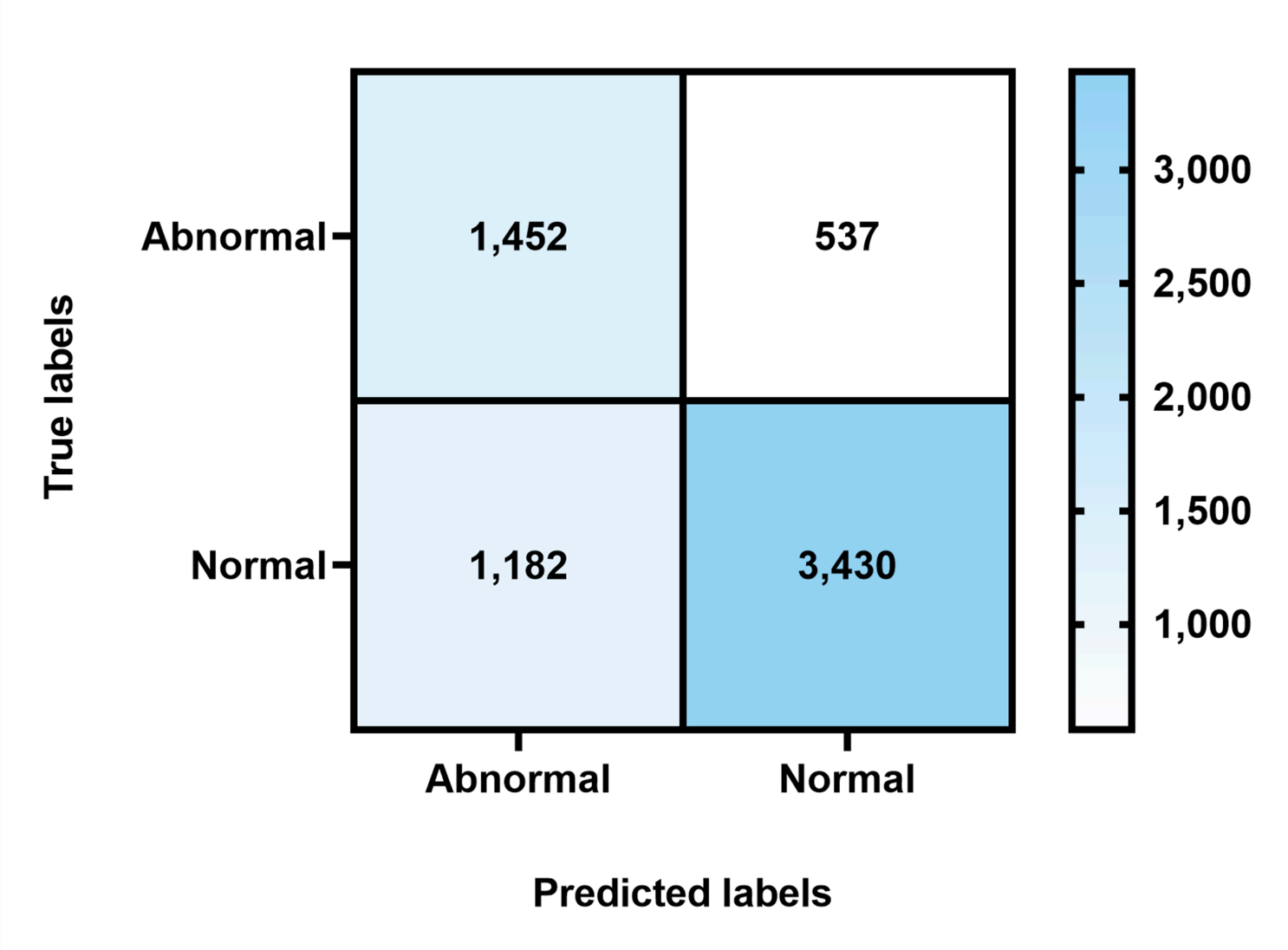


Figure 2

Confusion matrix illustrating the classification performance of the CNN model. The matrix shows the absolute number of slices labelled as abnormal or normal in the ground truth (y-axis) and predicted by the model (x-axis). Correct predictions appear along the main diagonal (abnormal/abnormal and normal/normal), while off-diagonal entries indicate misclassifications.

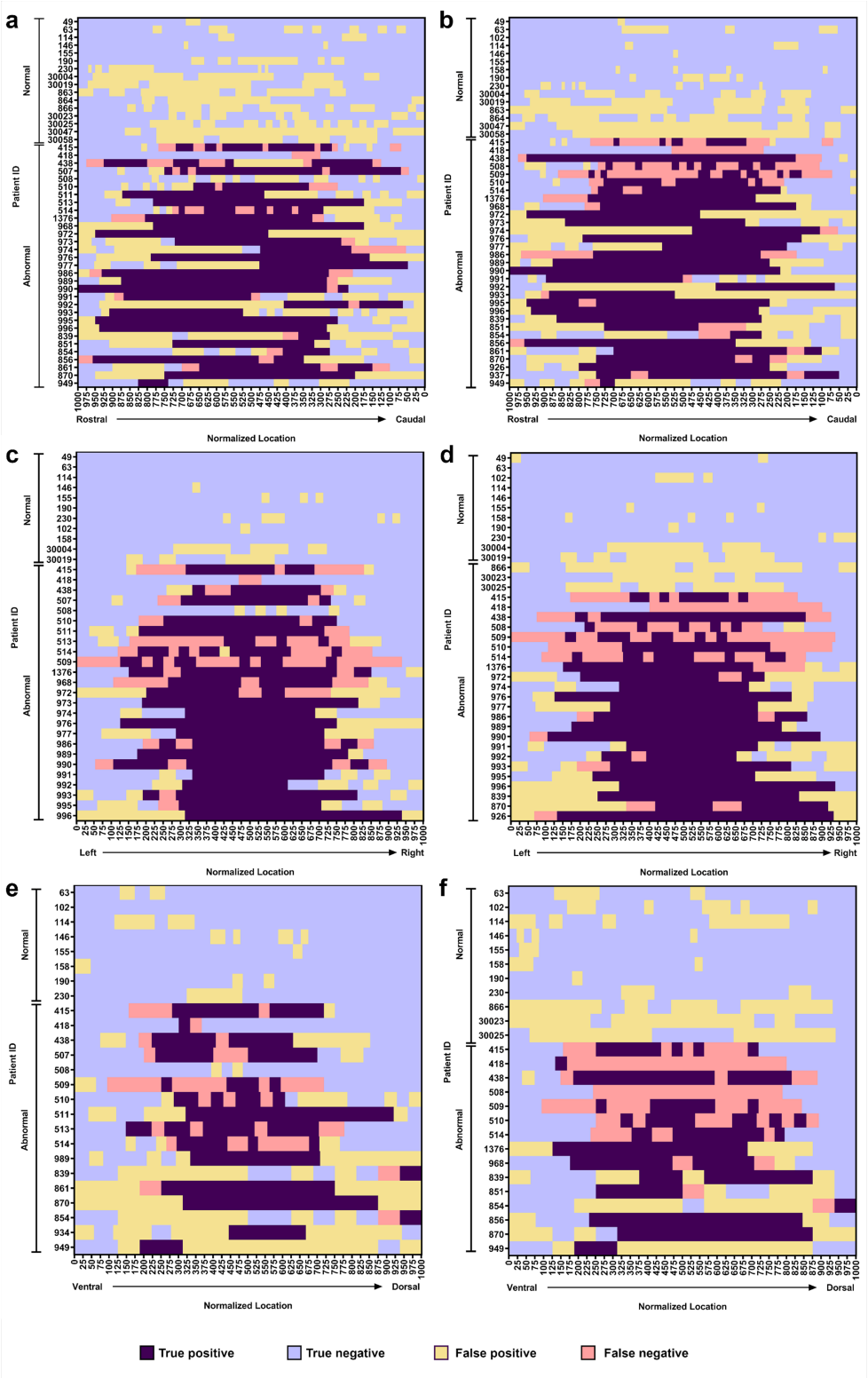


Figure 3

Spatial distribution of the slice-wise prediction outcomes after comparison with the ground truth across all six T1w MRI sequences in the test set, grouped by patient status (normal, abnormal). Each heatmap displays the classification outcome (true positive, violet; true negative, light blue; false positive, yellow; false negative, red) for individual slices along the normalized anatomical axis (x-axis). Panels represent: **(a)** transverse pre-contrast, **(b)** transverse post-contrast, **(c)** sagittal pre-contrast, **(d)**sagittal post-

contrast, **(e)** dorsal pre-contrast, and **(f)** dorsal post-contrast sequences. Normal patients are predominantly characterized by true negatives, whereas abnormal patients show dense clusters of true positives, especially in the central portion of the brain. The number of patients (y-axis) varies across the sequences due to differences in image acquisition.

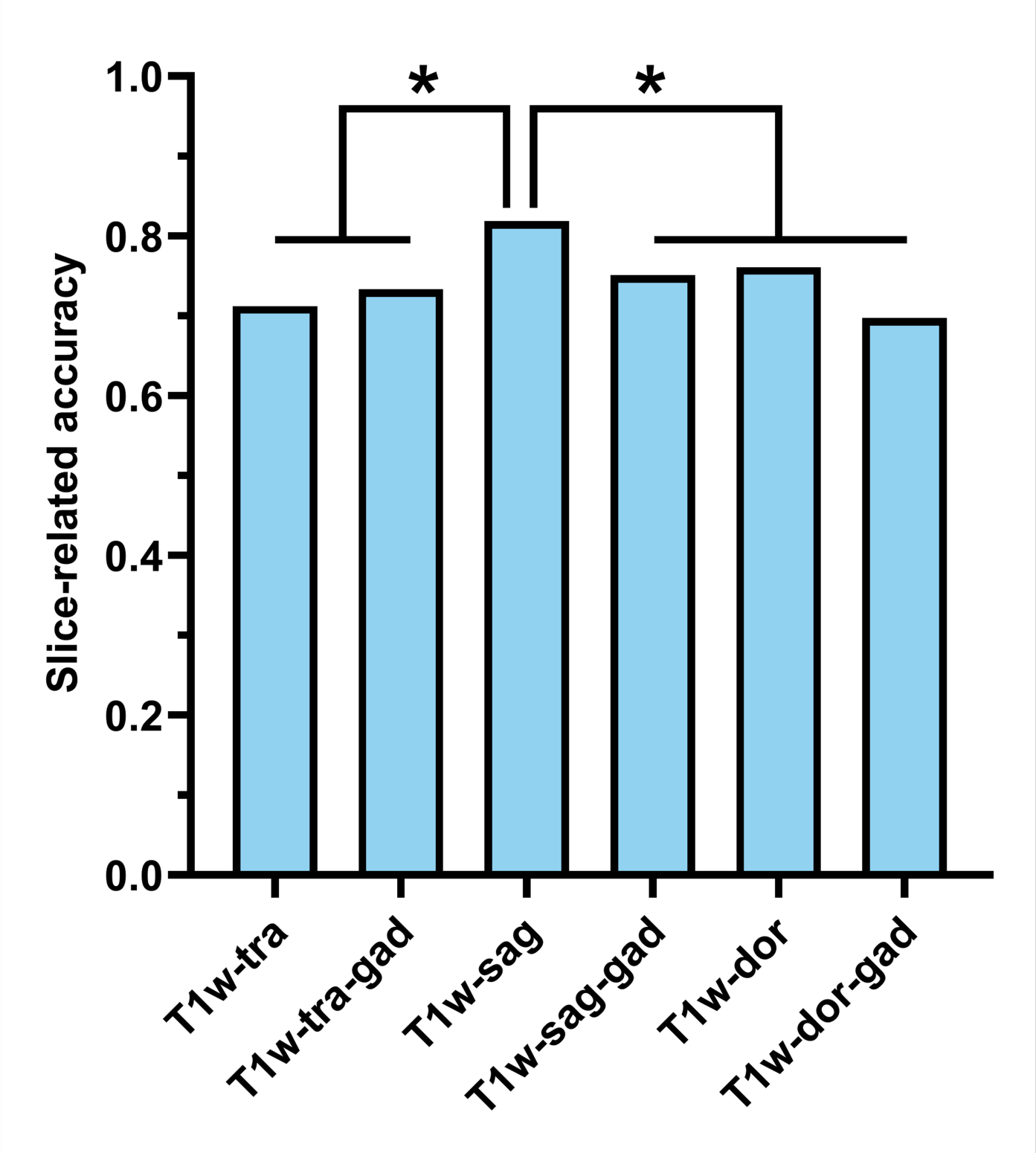


Figure 4

Comparison of slice-wise accuracy across T1-weighed MRI sequences. Overall accuracy differed significantly between the six sequences (Chi-square test, $p < 0.001$), with highest performance observed for sagittal pre-contrast images. Asterisks indicate statistically significant pairwise differences based on Fisher's exact tests with Bonferroni correction ($p \leq 0.05$).

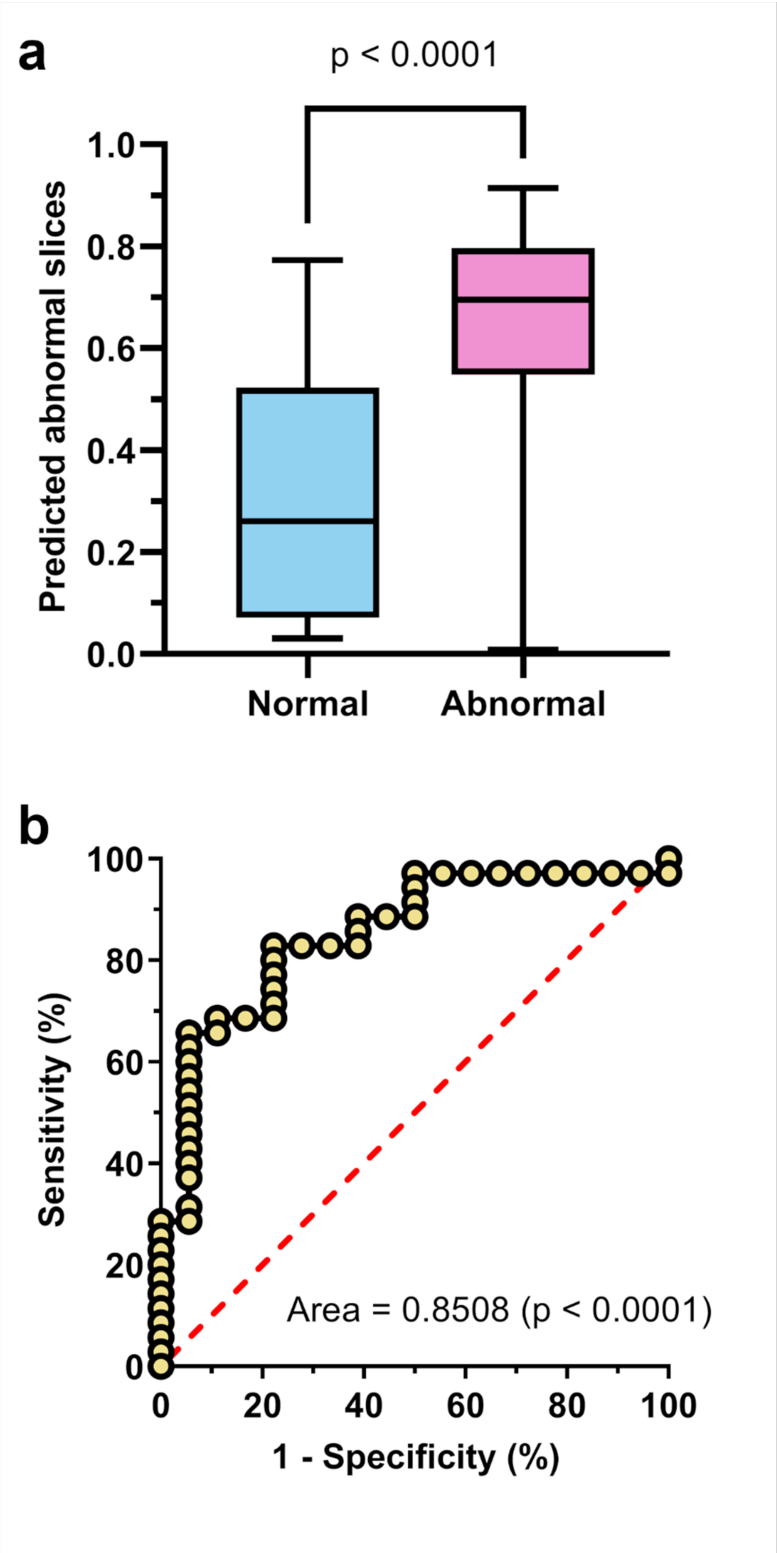


Figure 5

Performance of the CNN model for classifying patients as normal or abnormal based on the proportion of slices predicted as abnormal. **(a)** Box-and-whisker plot showing that the proportion of slices predicted as abnormal is significantly higher in clinically abnormal patients than in normal patients (Mann–Whitney U test, $p < 0.0001$). **(b)** Receiver operating characteristic (ROC) curve illustrating how a varying threshold of slices predicted as abnormal affects the sensitivity and specificity at the patient level. The area under the curve (AUC) is 0.8508 ($p < 0.0001$), indicating robust model performance.

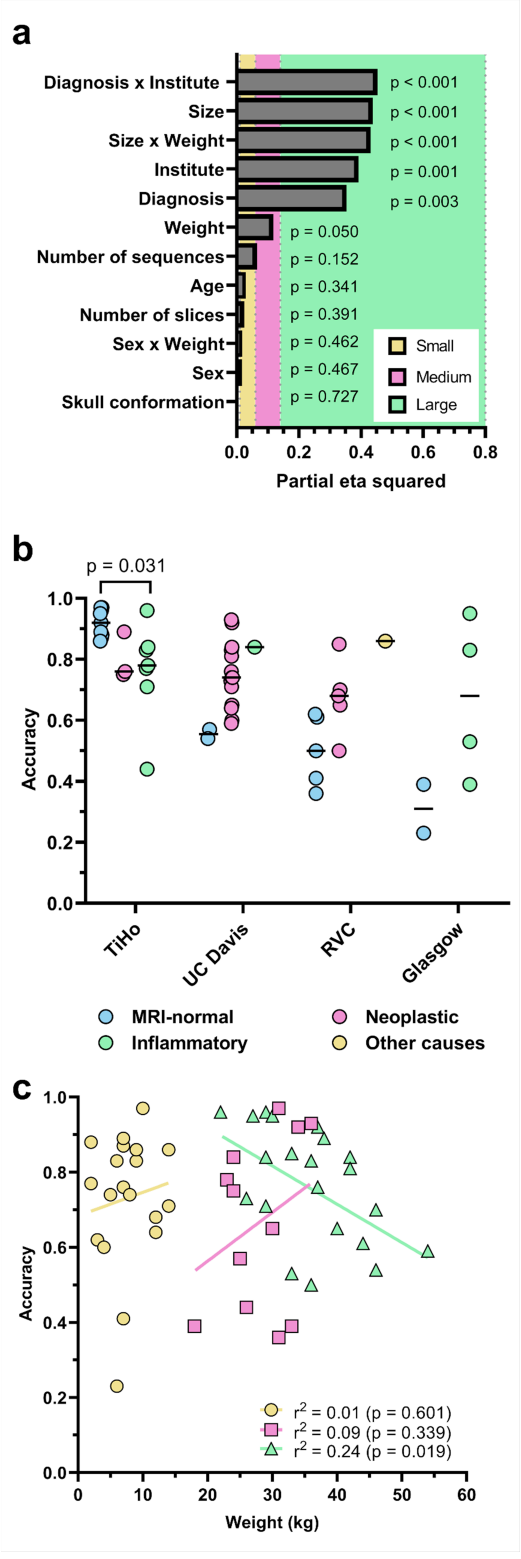


Figure 6

Analysis of factors influencing the classification accuracy of the CNN model. **(a)** The effect sizes from ANCOVA reveal that diagnosis and institute are the strongest predictors of model performance. **(b)** Dot plot of the accuracy values stratified by diagnosis and institute, assessed by two-way ANOVA. Post hoc comparisons of diagnoses within each institute were performed using one-way ANOVA with pairwise post hoc tests with Bonferroni correction. The asterisk indicates a statistically significant difference ($p < 0.05$). **(c)** Scatter plot of the accuracy values stratified by weight and breed size. Due to the significant interaction in the ANCOVA, the regression analysis was split for small (yellow), medium (red), and large (green) dog breeds. A significant negative effect of weight on the accuracy of the CNN model was observed only in large breed dogs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20250904SamiraAbaniSupplementaryMaterial.docx
- Sup.Figure1.tif
- Sup.Figure2.tif
- Sup.Figure3.tif
- Sup.Figure4.tif
- Sup.Figure5.tif
- Sup.Figure6.tif