



OPEN

APIS: a paired CT-MRI dataset for ischemic stroke segmentation - methods and challenges

Santiago Gómez¹, Edgar Rangel¹, Daniel Mantilla², Andrés Ortiz², Paul Camacho², Ezequiel de la Rosa^{3,4}, Joaquin Seia⁵, Jan S. Kirschke^{4,6}, Yihao Li^{7,8}, Mostafa El Habib Daho^{7,8} & Fabio Martínez¹✉

Stroke, the second leading cause of mortality globally, predominantly results from ischemic conditions. Immediate attention and diagnosis, related to the characterization of brain lesions, play a crucial role in patient prognosis. Standard stroke protocols include an initial evaluation from a non-contrast CT to discriminate between hemorrhage and ischemia. However, non-contrast CTs lack sensitivity in detecting subtle ischemic changes in this phase. Alternatively, diffusion-weighted MRI studies provide enhanced capabilities, yet are constrained by limited availability and higher costs. Hence, we idealize new approaches that integrate ADC stroke lesion findings into CT, to enhance the analysis and accelerate stroke patient management. This study details a public challenge where scientists applied top computational strategies to delineate stroke lesions on CT scans, utilizing paired ADC information. Also, it constitutes the first effort to build a paired dataset with NCCT and ADC studies of acute ischemic stroke patients. Submitted algorithms were validated with respect to the references of two expert radiologists. The best achieved Dice score was 0.2 over a test study with 36 patient studies. Despite all the teams employing specialized deep learning tools, results reveal limitations of computational approaches to support the segmentation of small lesions with heterogeneous density.

Keywords Ischemic stroke, Computed tomography, Image segmentation, Paired dataset, Deep learning

Stroke is the second leading cause of mortality worldwide and the most significant adult disability in developed countries¹. Ischemic stroke (IS), caused by blood vessel occlusion, is the most prevalent type of stroke, reporting 80% of all stroke cases². Brain tissue is extremely sensitive to ischemia, producing irreversible damage within minutes from the onset. Therefore, targeted therapies must be delivered within minutes to hours from symptoms onset for maximum effectiveness³.

Current clinical guidelines for stroke assessment establish that the first-line imaging modality is a non-contrast CT (NCCT), allowing support therapy and prognosis decisions, as long as there exists sufficient symptom, and clinical information^{4,5}. Among others, the NCCT may help to differentiate stroke mimics, allowing to exclude brain ischemia and hemorrhage, and being determinant to obtain a first approximation of arterial occlusion⁶. Considering immediate time attention, the NCCT is ideal due its speed, low-cost, and accessibility in clinical centers. Recent trials have confirmed the importance of well-defined therapeutic decisions, such as the use of endovascular thrombectomy in patients with large ischemic lesions visible on NCCT⁷. However, in some stroke cases, a complementary analysis from other imaging modalities is required, particularly when the stroke occurs during sleep or upon awakening^{8,9}. Also, the NCCT has low sensitivity in the early stages of ischemia, limiting the lesion characterization from observational expertise.

To overcome such limitations, neuroradiologists may observe lesions from diffusion-weighted imaging (DWI) sequences, such as DWI B-1000 and the apparent diffusion coefficient (ADC), which are biomarkers of cytotoxic injury that predict edema formation and outcome after IS¹⁰. However, this observational task is challenging, time-consuming (can take up to 15 minutes per case depending on the clinical workflow and expertise), and

¹Biomedical Imaging, Vision, and Learning Laboratory (BIVL2ab), Universidad Industrial de Santander, Bucaramanga, Colombia. ²Clínica FOSCAL, Floridablanca, Colombia. ³icometrix, Leuven, Belgium. ⁴Department of Informatics, Technical University Munich, Munich, Germany. ⁵University of Girona, Girona, Spain. ⁶Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, University of Munich, Munich, Germany. ⁷LaTIM UMR 1101, Inserm, Brest, France. ⁸University of Western Brittany, Brest, France. ✉email: famarc@uis.edu.co

susceptible to bias errors^{11,12}. Additionally, ischemic lesion delineation on NCCT reveals lower inter-rater and spatial agreement than on DWI.

Recent literature has notably highlighted the increased use of technological solutions to characterize lesions from various imaging modalities, thereby supporting different steps in diagnosis and prognosis. The promise of these tools lies in the capability to stratify patient subgroups to carry out personalized and more detailed analyses. Technological alternatives and commercial tools developed from NCCT images have enhanced stroke care by increasing the speed and standardization of attention, as well as improving the quantitative accuracy of boundary lesion definition^{13,14}. Particularly, Cina-ASPECTS, e-ASPECTS, and Rapid ASPECTS complement standard evaluation of early ischemic changes in the middle cerebral artery territory, offering complementary information to bring rapid assessment and reducing subjective interobserver variability^{15,16}. Other commercial solutions such as DeepCT, Stroke VCAR, CuraRad-ICH, Viz ICH have been designed to support the detection and coarsely localize hemorrhage-associated findings. Preliminary results evidence that these tools reduce patient attention time. However, conclusive evaluations are lacking due to the absence of rigorous statistical validations in experiments closely resembling clinical practice^{17,18}. Also, for computational solutions the low sensitivity is challenging to design generalizable solutions that may interact among images from different scanners, and patient groups with marked variability⁶. Importantly, these studies also report limitations of such alternatives, standing out NCCT cases where the proposed method may fail¹³.

Hence, the characterization of lesions remains an open problem due to the complexity of low-contrast observations and the lack of CT paired with other modalities to guide segmentation during training¹⁹. Despite multiple efforts, a main drawback within the scientific community remains the scarcity of annotated stroke data to formulate and adjust computational approaches for supporting lesion segmentation in NCCT studies. Particularly, the paired DWI-NCCT training may help in the early detection of radiological findings, allowing rapid decisions on the appropriateness of interventional treatments (i.e., mechanical thrombectomy or thrombolysis) for stroke patients⁷. These efforts may also support additional transparent validations of artificial intelligence methodologies, allowing bringing real evidence about these approximations into clinical practice. In the same line, these developed tools may guide and help less experienced doctors during the process of interpreting ischemic stroke CT scans.

This work reports the experience of a stroke segmentation challenge, where scientific participants take top computational architectures to segment CT studies, considering a publicly available dataset featuring paired CT-MRI scans of acute IS patients, with annotations from two expert radiologists. Participants were invited to propose computational approaches that take advantage of paired information to carry out delineations over CT studies. As for the testing stage, we carried out a hidden evaluation of participant strategies using NCCT scans with delineations from two expert radiologists. The evaluation included the comprehensive assessment of the submitted algorithms with overlapping, clinical, and boundary metrics. Additionally, this study also included an agreement analysis between two expert radiologists regarding lesion delineations, thereby establishing a baseline for segmentation approaches in NCCT lesion detection and ADC delineation. The evaluation of proposed strategies against studies with varying stroke patterns and lesion sizes highlighted the challenges in segmenting IS lesions from NCCT. Specifically, it exposed the limitations of computational methods in segmenting lesions that were also undetected by radiologists on NCCT, particularly smaller lesions.

Current works

Stroke data

Since 2015, the notable increase in the availability of public datasets for stroke have been crucial for the advancement and validation of computational algorithms, supporting stroke segmentation. A significant amount of research has been directed towards MRI datasets for IS patterns detection^{20,21}, with alternative diffusion studies^{22–25}.

Some CT initiatives include the Acute Ischemic Stroke Dataset (AISD) dataset²⁶ with 397 CT-MRI pairs. CTs were obtained within 24 h following symptom onset, with subsequent DWI imaging conducted within the next 24 h. Notably, these annotations exhibit lesion volumes with a median (IQR) of 19.26 (8.25–46.82) mL, evidencing a lack of small lesions. This can be attributed to the imaging timeline, where the extended period until the capture of the MRI scan allowed it to evolve lesions and become more pronounced, resulting in larger observed lesion volumes. Also, one critical aspect of the AISD dataset is the availability of test samples, giving participants the opportunity to overfit their strategies. On the other hand, the Ischemic Stroke Lesion Segmentation dataset of 2018 (ISLES2018)²⁷ provides CTP parametric maps such as cerebral blood flow (CBF), cerebral blood volume (CBV), mean transit time (MTT), and time to maximum (TMax). Participants were recruited from two prospective IS trials^{28,29}, identified with acute large artery occlusive IS and absence of hemorrhage within 8 h of the last known normal state. Manual annotations of ischemic lesions were carried out by one stroke neurologist from DWI images, acquired within 3 h of CTP. The median (IQR) lesion volume of the patients in the train split was 16.57 (7.23–44.14) mL, indicating lesion volumes larger than those typically seen in early subacute stroke stages. Nevertheless, such DWIs were only available at early stages of the challenge and cannot be included in the design of new segmentation strategies. Additionally, while CTP studies offer insights into cerebral blood flow and volume, the determination of the arterial input function (AIF) is susceptible to bias in selecting appropriate values³⁰.

Stroke strategies to support segmentation

Computational strategies from the AISD dataset²⁶ have leveraged the anatomical symmetry information to exploit the pixel-wise difference between hemispheres^{31,32}. Also, unsupervised nets have been proposed to obtain bilateral aligned images, standing out in-planar and across-planar symmetry²⁶. Likewise, Ni et al.³³ proposed three 3D CNNs to sequentially align NCCT inputs across the mid-sagittal axis. Complementary, Xu et al.³⁴

implemented a residual convolutional autoencoder with six deformable transformer layers in the bottleneck to segment ischemic lesions from 3D NCCT inputs. These approaches, nonetheless, remain limited due to the observed low signal-to-noise and contrast-to-noise ratios of brain tissues, preventing computational nets from learning more sophisticated descriptors.

From ISLES2018 data²⁷, some approaches used a 3D U-Net based model with dilated convolutions³⁵, while other strategies proposed take image modalities separately with multiple dilated convolutional encoders³⁶. These representations remain limited to recover whole context information due to error propagation on skip connections. Also, Cl  rigues et al. proposed a 2D asymmetric residual encoder-decoder framework, following a dynamically weighted loss function and mini-batches of CT and CTP patches³⁷. The use of CTP maps allowed computational approaches to improve their lesion segmentation capabilities. Nevertheless, the integration of tissue density information from CT and hemodynamic data from CTP presents significant challenges. Specifically, artifacts and variability in the CTP data, resulting from inadequate selection of AIF and VOF, may be propagated into the joint representation, potentially impacting the overall analysis³⁰. Nonetheless, strategies following synthetic DWI approximations remain as the top performers in this dataset, mainly because of the availability of the rich paired information in the initial version of the dataset. For instance, Liu et al.³⁸ proposed an adversarial-generation of synthetic MRI-DWI images from CTP studies, mapped through a segmentation network to predict the lesion outcome. Similarly, Wang et al.³⁹ proposed a pipeline to generate synthetic DWI sequences from 4D Perfusion CT sequences. Despite their remarkable performance, these studies cannot be reproduced because paired and aligned DWIs are currently not publicly available. Moreover, obtaining such paired CT-MRI studies within a short-time window is challenging in clinical scenarios. Isolated efforts over such translation schemes have been reported between FLAIR and NCCT images, but following owner datasets⁴⁰.

Other computational alternatives have used private databases with relevant clinical variables and scales, captured during the stroke patient management, such as the ASPECTS score⁴¹. Furthermore, Gauriau *et al.* implemented a classical UNet to segment acute ischemic lesions from 3556 NCCTs studies⁴², but remains constrained in terms of replicating results, developing new solutions, and validating its ability to generalize.

Methods

APIS data

In this work is introduced a paired NCCT-ADC dataset, carefully built to exploit complementary radiological findings and support stroke lesion segmentation. This dataset was initially presented in the ISBI official challenge “APIS: A Paired CT-MRI Dataset for Ischemic Stroke Segmentation Challenge”. To build the dataset, a retrospective study was conducted to validate collected 96 studies of patients presenting with stroke symptoms at two clinical centers between October 2021 and September 2022. The studies were categorized into control ($n = 10$) or IS ($n = 86$) studies. Control patients with stroke symptoms were included to diversify tissue samples, potentially enhancing deep learning models’ ability to segment stroke lesion tissue. For each study, the triage NCCT and the subsequent ADC were acquired. The inclusion criteria involve patients older than 18 years, without signs of cerebral hemorrhage, with no treatment between NCCT and MRI image acquisition, and confirmed IS diagnosis after reviewing imaging studies. Also, if there was evidence of partial reperfusion, patients were not excluded.

Figure 1 summarizes the retrospective protocol to collect data, including the filtering process and demographic data in the train and test sets. The studies were captured using two CT scanners: (i) Toshiba Aquilion 64 TSX-101 and (ii) Toshiba Aquilion one TSX-301/ac scanner, with 64 and 320 detectors, respectively. Also, two MRI scanners, namely (i) Toshiba Vantage Titan MRI and (ii) General Electric SIGNA Explorer, both with a magnetic field of 1.5 T. The ischemic lesion annotations were carried out by two neuro-interventional radiologists with more than five years of experience. Firstly, experts reviewed the clinical record at admission, and subsequently, identified whether the lesion was visible in the NCCT modality. The clinical records included demographic details, and symptom identification (such as unilateral weakness or paralysis, speech difficulties, vision disturbances, confusion or alterations in consciousness, severe headache, and dizziness or balance problems). Also, the patient’s medical history, including hypertension, diabetes, and dyslipidemia. Additionally, for

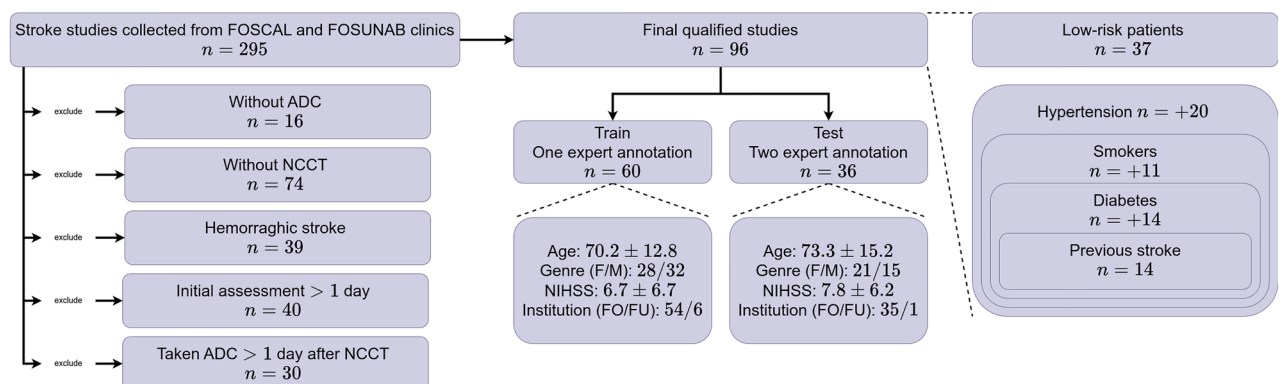


Fig. 1. Participant selection and demographic distribution for the training and testing partitions of the APIS dataset.

some cases, it was included the assessment of severity and functionality, the duration between symptom onset and imaging, and information on the patient's response to previous treatments and any medical complications to formulate effective treatment plans. Finally, experts assessed DWI and ADC images and carried out the manual delineations on ADC. The lesion annotation process was performed by each radiologist individually using the MRICroGL software⁴³. All studies were skull-stripped using optiBET⁴⁴ and co-registered having as reference the labeled ADC sequences with Elastix⁴⁵. Figure 2 shows a preview of the image modalities and the respective segmentation mask. The delineated lesions show variability in shape, size, and number of consecutive slices where the lesion occurs. The annotated dataset remains accessible to the public upon registration, inviting the scientific community to deal with stroke characterization from NCCT but guided with paired DWI information.

Ethics statement

The research was performed in accordance with relevant guidelines/regulations of the Declaration of Helsinki. To ensure complete anonymity, all patient information was eliminated from the volumetric nifty files, and any facial features were removed through skull stripping. The experimental protocol was approved by the ethics committees *CEINCI-UIS* at Industrial University of Santander (approval number 4110, granted on February 10, 2023) and *CEI-FOSCAL* at Ophthalmological Foundation of Santander - FOSCAL (approval number 06075/2022, granted on April 29, 2022), both located in Colombia. Due to the retrospective nature of the study, ethics committees *CEINCI-UIS* at Industrial University of Santander and *CEI-FOSCAL* at Ophthalmological Foundation of Santander—FOSCAL waived the need of obtaining informed consent.

The evaluation and ranking system

Participants were invited to submit their algorithms to predict stroke lesions on a hidden test set. For APIS evaluation, the segmentations with expert delineations, take into account overlapping, clinical, and boundary metrics. In terms of overlapping metrics, the Dice score (DICE), precision (PREC), and sensitivity (SENS) were included. To comprehensively analyze the evaluated approaches in clinical practice, the study considered metrics that elucidate important lesion characteristics. Consequently, this study introduced as clinical metrics the absolute lesion volume difference (AVD), absolute lesion difference (ALD), and lesion detection F1 score (LF1). The AVD metric highlights deviations in model-predicted lesion volume, which is critical for treatment planning and prognosis. The ALD and LF1 metrics assess annotation granularity, aiding in stroke etiology determination. To assess lesion boundary definition, the Hausdorff Distance (HD) using an Euclidean distance was adapted, considering special cases where there are no stroke annotations from the proposed models or from the radiologist. Particularly, in the studies that lacked a comparative annotation (from a radiologist reference or an output model segmentation), the maximum distance between points of the available stroke annotation and the brain masks was computed. In the same line, for control cases where models erroneously predicted lesions, the Euclidean distance from the most distant points in the brain parenchyma was assigned to maintain consistent penalization.

Regarding the ranking methodology, the APIS challenge adopted the “rank then aggregate” methodology⁴⁶ followed in previous academic competitions in IS segmentation^{21,24}. All clinical and boundary metrics were considered in the ranking systems while overlapping only considered the DICE score. For every patient in the test split, metric-specific rankings were computed following an average ranking methodology to provide a fair and balanced representation for statistical purposes. Individual patient case rankings were determined from the mean of all metric rankings, while radiologist-specific rankings were derived by averaging the metric rankings pertinent to each respective rater. Conclusive rankings were derived by averaging case rankings following a

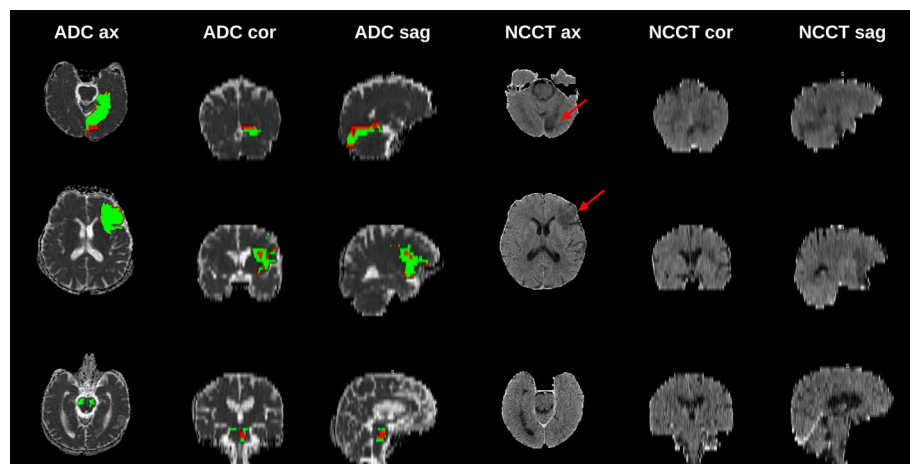


Fig. 2. Examples of images in the APIS dataset. ADC images in axial, coronal, and sagittal planes are displayed in the leftmost columns, while matching views of NCCT images appear in the rightmost columns, both displaying annotations of two experts. The top rows highlight lesions manifesting as hypointense on ADC and hypoattenuated on NCCT. The bottom row presents a case where the lesion is discernible on ADC but less visible on NCCT.

quantitative measure of each participant's predictions alignment with expert delineations. For the final ranking report, a rank of one represented the optimal performance, while a rank of four denoted the least favorable outcomes in comparison to peers.

Strategies from APIS participants

From the APIS challenge, participants were invited to propose computational approaches that take advantage of paired CT-ADC studies to carry out delineations over CT studies. The APIS challenge opened in December 2022 and received inscriptions from a total of 41 teams worldwide. Training data was released on December 26th of 2022, and over the next four months, teams had the opportunity to design and implement strategies for segmenting acute IS lesions from the NCCT images, supported by paired ADC sequences. All teams had access to annotations from only one radiologist. For the test phase, the four finalist teams were: *icomaia*, *XMU*, *LONGYING*, and *LaTIM*. The solutions proposed by the four finalists are summarized below.

Icomaia solution

The *icomaia* solution was developed to exploit inter-hemispheric tissue differences through a process involving image preprocessing, model selection, and training strategy optimization. Regarding image preprocessing, potential misalignments between NCCT and MRI images were rectified through cross-modality registration. For registration, all images underwent resampling to a uniform voxel resolution of $1\text{mm}\text{\AA}^3$. To emphasize structural information for skull-stripping, a pseudo-T1 image was generated from the resampled NCCT with SynthSR⁴⁷. Subsequently, the pseudo-T1 underwent skull-stripping with SynthStrip⁴⁸, and the resultant brain mask was applied to the resampled NCCT. The stroke lesion masks were propagated to the NCCT space using the transformation parameters obtained from the registration of the ADC to the skull-stripped NCCT with Elastix⁴⁵. To focus on the inter-hemispheric differences, a contralateral NCCT difference image was computed, which involved registering the image to an NCCT MNI space template⁴⁹, left-right flipping, and come back projection to the original NCCT image space for the final subtraction. The last preprocessing step involved NCCT intensity clipping within the range $[-100, 400]$ HU. For the model selection, self-configuring nnU-Net with default parameters were considered following 2D and 3D full-resolution versions⁵⁰. The model inputs are pre-processed NCCT and the contralateral difference image in a bi-channel configuration. Models adjustment consisted of a pre-training stage using the AISD dataset²⁶, followed by fine-tuning on the APIS samples. Each training lasted one hundred epochs under a stratified 70-20-10% training-validation-test scheme, ensuring equal representation of right, left, and bilateral lesions, as well as similar lesion sizes. Pre-training provided the models with a wider variety of stroke samples, thereby enhancing their performance and ensuring faster convergence during subsequent fine-tuning on the APIS dataset. The 3D full-resolution model obtained the best results, thus serving as the final challenge submission.

XMU solution

The *XMU* team optimized the nnU-Net⁵⁰ for small lesion sizes in the APIS dataset, training the model exclusively on NCCT images. The *XMU* team optimized the nnU-Net⁵⁰ for small lesion sizes, training the model exclusively on NCCT images. To standardize and mitigate outliers' impact in the training process, each NCCT image was resampled using a third-order spline interpolation, clipped to the intensities between the 0.5 and 99.5 percentiles, and normalized by the subtraction and division of the mean and standard deviation of the brain voxels, respectively. Subsequently, all cases were resampled to the median spacing of the training cases. Regarding the model, given the small size of the lesions in the APIS dataset, a 2D U-Net was configured to receive NCCT patches of 128×128 pixels. During training, composite Dice and cross-entropy loss were used to optimize for global overlap and fine-grained pixel-wise optimization. Additionally, to account for the lesion's size and shape, the loss signal was computed for each instance prediction. The model was trained over 1000 epochs using an SGD optimizer with Nesterov momentum (0.99) with a learning rate of 0.01 with poly reduction.

LONGYING solution

The *LONGYING* team proposed an ensemble of three convolutional and transformer autoencoders to extract a diverse set of features from the NCCT images. More precisely, to consider the fine-grained details of slices and volumetric information of stroke lesions, 2D (efficientnet-B6⁵¹ and CoAtNet⁵²) and 3D backbones (ResNet⁵³) were implemented. Each model was trained separately following a stratified group 5-fold cross-validation and a balanced sampler to ensure a fair data representation during training runs. The training was carried out using an SGD optimizer, and a focal loss⁵⁴ to deal with class imbalance and small lesions. To address the weight oscillation problem of traditional SGD in the reverse process, at the end of every epoch a stochastic weight averaging (SWA) was applied to stabilize the model parameters and improve generalization performance. As for the data preprocessing, NCCT images underwent a CLAHE normalization, and axial slices were extracted from the normalized volumes for the 2D models. Data augmentation methods, such as horizontal and vertical flips and rotations, were employed to strengthen the models against overfitting. The final predictions were obtained following an average of the three architecture predictions. Test-time augmentations were employed to improve the reliability and accuracy of predictions by enhancing robustness and stability through the aggregation of multiple perspectives.

LaTIM solution

The *LaTIM* team proposed a multimodal segmentation strategy utilizing NCCT and synthetic ADC maps, generated through an image translation method. Specifically, this team proposed the adjusting of a 3D Pix2Pix Generative Adversarial Network (GAN)⁵⁵. This solution is used as the backbone to represent NCCT features the UNETR⁵⁶, thus, generating synthetic ADC maps from NCCT sequences. For the discriminator model, a simple

CNN architecture was used. Subsequently, an early fusion was applied to integrate NCCT and synthetic ADC maps as separate channels in the input of a multimodal U-Net. The training protocol incorporated data augmentation techniques, such as elastic transformations, vertical and horizontal flips, and the addition of Gaussian noise, with the Adam optimizer. The generative model applied a combination of binary cross-entropy and L1 losses (the latter multiplied by 100), whereas the discriminative model exclusively used binary cross-entropy loss. Given the complexity of generative models, they underwent training once for every five batches. The dataset was partitioned into 55 samples for training and 5 for validation. For preprocessing, NCCT images were resized to 26 slices of 256×256 pixels to ensure uniformity in the training data. For evaluation, the generative model first estimated the synthetic ADC map. Subsequently, the synthetic ADC and the real NCCT were then concatenated in the channel dimension and fed into the segmentation model to generate the final stroke lesion prediction.

Evaluation and results

Agreement from radiologist annotations

An agreement study between expert lesion references established challenges associated with the characterization of findings associated with stroke lesions. In this study, in a first step, the agreement analysis was carried out between two expert neuroradiologists. This study included an agreement analysis of the expert annotations on i) lesion detection from NCCT, and ii) stroke lesion delineation from ADC maps.

Regarding lesion detection, there was a consensus among the radiologists regarding the presence or absence of a stroke lesion in 31 of 36 test cases. This consensus, measured using the Cohen's kappa coefficient, showed a moderate agreement ($\kappa=0.72$) for lesion identification from NCCT scans. Experts concurred on lesion visibility (*BCT+*) in thirteen instances (36.1%) and agreed on lesion absence (*BCT-*) in 18 cases (50%). These findings highlight a considerable number of NCCTs in which the injury is not visible for experts. Moreover, there were five instances where the radiologists did not reach a consensus. Notably, in four cases (11.1%), only the first radiologist identified a lesion (*R1CT+*), while in one case (2.8%), only the second radiologist identified a lesion (*R2CT+*).

Concerning the delineations made on ADC, a mean Dice score of 0.63 ± 0.29 and volume difference of 4.45 ± 7.24 mL were calculated, underscoring the significant variations in the lesion delineation process among experts. Figure 3 shows the histograms of lesions volumes for each radiologist. An analysis of lesion volumes showed that the median (IQR) volume of the lesion annotations by R1 was 4.3 (0.57–31.31) mL and 8.44 (1.19–42.38) mL for R2, indicating a prevalence of small lesion volumes. Additionally, it can be noted that R2 delineated slightly larger volumes than R1. We further refer to lesions below 1.19 mL as *tiny*, between 1.19 and 8.44 mL as *small*, between 8.44 and 42.38 mL as *medium* and lesions exceeding 42.38 mL as *large*. Notably, within the *BCT+* group, median volumes were 18.60 (6.02–56.56) mL for R1 and 20.00 (12.64–57.36) mL for R2, reflective of a tendency towards *medium* sized lesions. Within the *R1CT+* group, the median lesion volumes were noted as 50.19 (3.32–99.85) mL for R1 and 51.00 (4.97–98.94) mL for R2, which denotes a predominance of *large*-sized lesions in this group. On the other hand, for the single case in the *R2CT+* group, experts delineated lesions classified as *tiny*, with volumes measured at 0.60 mL for R1 and 1.20 mL for R2. Conversely, the *BCT-* group presented with smaller median lesion volumes, reported at 3.37 (0.63–10.34) mL for R1 and 3.76 (0.94–11.97) mL for R2, highlighting the predominance of *small* lesion volumes in these cases. Moreover, the two radiologists classified three cases as control (*BCT-Con*). Surprisingly, in one case the second radiologist annotated a small lesion (7.29 mL), while the first classified the case as control (*R1CT-Con*). As an additional study to improve the statistical robustness of the agreement study and to corroborate the challenges associated with stroke lesion delineation, an extra neuroradiologist with over five years of experience was included to annotate the test cases following the same protocol. The lesion delineations exhibited volumes of 4.28 (0.69–30.82) mL, comparable to those of R1. Furthermore, the agreement levels remained consistent when comparing R1 and the additional neuroradiologist, with a Dice score of 0.66 ± 0.29 and an AVD of 4.77 ± 10.17 mL. However, the metrics for the comparison with R2 annotations were slightly lower, with a Dice score of 0.57 ± 0.29 and an AVD of 6.93 ± 10.05 mL.

The *BCT+* category exhibited the highest Dice score (0.69 ± 0.22) and the third best volume difference (7.14 ± 10.55 mL). High variability in Dice and AVD was observed among groups, which may be a consequence of the limited sample size in some groups. The lowest volume difference was achieved in the *R2CT+* group, where the radiologists annotations showed a 0.60mL difference and a Dice score of 0.55. Interestingly, the second lowest volume difference was recorded on the *BCT-* with 2.12 ± 2.43 mL with a Dice score of 0.57 ± 0.29 , showing that challenges in identifying stroke lesions from NCCT, are not present in ADC. The *R1CT+* group achieved a

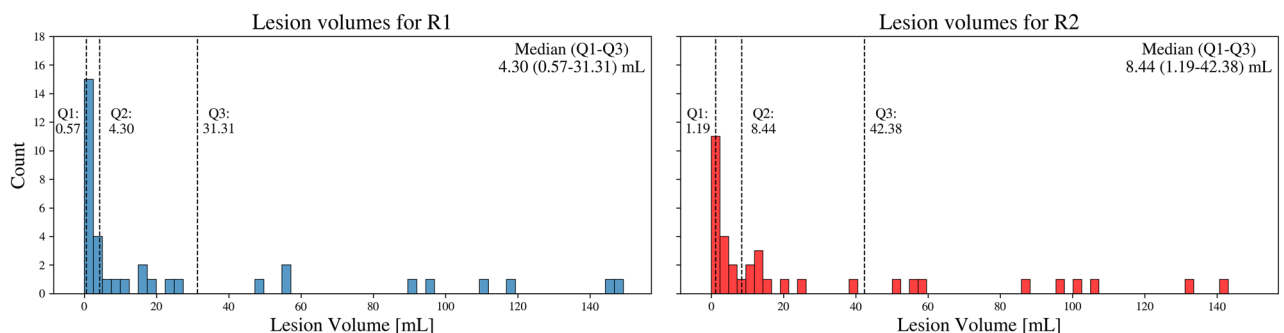


Fig. 3. Histograms of lesion volumes for the annotations of both radiologists over the cases in the test set.

slightly lower Dice score of 0.60 ± 0.33 in comparison to *BCT+* group, alongside a volume difference of 7.46 ± 6.63 mL. Perfect performance was noted in the *BCT-Con* category, with a Dice score of 1.0 and no volume difference, while *RICT-Con* presented a Dice score of 0.0 but a volume difference of 7.29mL due to a discrepancy in control case classification.

Results from participants

Table 1 presents the leaderboard that provides a detailed evaluation across participating teams. The *icomaia* team achieved the leading position based on quantitative assessments across specified metrics, with higher overlap and better defined lesion boundaries. Furthermore, *XMU* demonstrated consistency by achieving the second-best rankings across all metrics, ensuring a balance between classical segmentation and the requirements of clinical practice, which are closely related with accurate diagnostics and effective treatment.

Table 2 summarizes the best-achieved results per team regarding the annotations of both radiologists. The *icomaia* team localized the most number of ischemic lesions ($R1=17$ and $R2=20$). Contrarily, *XMU*, *LONGYING*, and *LaTIM* teams obtained sensitivities close to zero, indicating a low capability to identify lesions voxels on NCCT. For lesion boundaries estimation, *icomaia* obtained the best HD for both radiologists, suggesting that their strategy estimates better lesion boundaries. *XMU*, *LONGYING*, and *LaTIM* recorded HD scores exceeding 100 for both raters. As for the medical metrics, the best AVD scores were obtained by team *XMU*, while the *icomaia* team ranked last in this metric for both radiologists. For ALD, the *XMU* team obtained the best result, and team *icomaia* also ranked last for this medical metric, obtaining the highest lesion differences for both experts. Also, it is important to note that lesion count differences are considerably higher for R2 for all teams with the exception of *icomaia*, suggesting that R2 is more likely to annotate one volume with multiple segments. Considering the lesion-wise F1-score, the best score w.r.t. R1 annotations was attained by *XMU*. For R2, both *icomaia* and *XMU* demonstrated comparable performances. Examples of segmentations for all teams can be observed in Supplementary Fig. S1.

Table 3 summarizes the Dice scores and volume differences for NCCT groups. *icomaia* reported better performance in most groups with stroke lesions (*BCT+*, *RICT+*, *BCT-*). In *BCT+* were observed Dice scores of 0.38 for R1 and R2, and volume differences ($R1=42.66$ mL and $R2=37.86$ mL for R2). The lowest volume difference for R1 was reported by *XMU*. For the *BCT-* group, the Dice scores were lower across all teams, with *icomaia* obtaining the highest scores for both radiologists, indicating a challenge in lesion identification in these instances. However, the volume differences were more favorable for *LONGYING*, which achieved the lowest differences for both radiologists. In other groups with fewer samples, high variability was observed. It should be taking into account that for these groups with few samples there is not a conclusive statistical remark. Particularly, regarding the *RICT+* group, *icomaia* reported better Dice scores and volume differences concerning both radiologists. In the singular *R2CT+* case, *icomaia* was the only team to estimate a lesion, which resulted in no overlap for both radiologists. Regarding the volume difference, attributable to the tiny size of the lesion, *icomaia* showed larger discrepancies, indicative of an overestimation of the lesion's extent. In the *BCT-Con* group, all teams except *icomaia* accurately classified control subjects, achieving perfect dice and volume differences. The single case in *RICT-Con* exhibited variations due to the disagreement between the radiologists.

Place	Team	DICE Rank	HD Rank	AVD Rank	ALD Rank	LF1 Rank	R1 Rank	R2 Rank	Final Rank
1	<i>icomaia</i>	1.88	1.56	2.88	3.16	2.09	2.26	2.09	2.17
2	<i>XMU</i>	2.58	2.68	2.35	2.27	2.47	2.48	2.51	2.49
3	<i>LONGYING</i>	2.74	2.75	2.34	2.22	2.70	2.57	2.63	2.60
4	<i>LaTIM</i>	2.81	3.01	2.45	2.37	2.76	2.69	2.77	2.73

Table 1. APIS challenge leaderboard based on combined expert evaluations. The best values are bolded, and the second-best values are italicized. Lower ranks are better.

	Team	Intersect	DICE \uparrow	SENS \uparrow	PREC \uparrow	HD \downarrow	AVD \downarrow	ALD \downarrow	LF1 \uparrow
R1	<i>icomaia</i>	17/36	0.20 \pm 0.26	0.27 \pm 0.32	0.18 \pm 0.25	79.57 \pm 40.58	32.41 \pm 40.02	6.81 \pm 4.15	0.17 \pm 0.24
	<i>XMU</i>	10/36	0.13 \pm 0.32	0.01 \pm 0.04	0.14 \pm 0.32	104.83 \pm 56.19	26.97 \pm 41.48	1.36 \pm 1.22	0.24 \pm 0.41
	<i>LONGYING</i>	7/36	0.12 \pm 0.32	0.00 \pm 0.01	0.05 \pm 0.19	107.29 \pm 58.43	27.59 \pm 43.94	1.47 \pm 1.67	0.15 \pm 0.33
	<i>LaTIM</i>	6/36	0.11 \pm 0.32	0.00 \pm 0.00	0.00 \pm 0.01	120.59 \pm 55.99	27.66 \pm 44.69	1.42 \pm 1.50	0.16 \pm 0.36
R2	<i>icomaia</i>	20/36	0.20 \pm 0.24	0.29 \pm 0.32	0.17 \pm 0.24	73.95 \pm 38.55	30.69 \pm 37.91	6.14 \pm 3.97	0.21 \pm 0.24
	<i>XMU</i>	10/36	0.10 \pm 0.28	0.01 \pm 0.05	0.15 \pm 0.34	117.07 \pm 58.93	27.20 \pm 38.73	3.03 \pm 2.90	0.21 \pm 0.38
	<i>LONGYING</i>	6/36	0.09 \pm 0.28	0.00 \pm 0.01	0.04 \pm 0.18	118.25 \pm 53.35	27.81 \pm 40.98	2.86 \pm 2.92	0.11 \pm 0.30
	<i>LaTIM</i>	4/36	0.08 \pm 0.28	0.00 \pm 0.00	0.00 \pm 0.02	129.83 \pm 53.11	27.94 \pm 41.60	3.08 \pm 3.05	0.11 \pm 0.32

Table 2. Summary of the NCCT segmentation metrics performance for the top four teams regarding the criteria both experts. The number in the check-mark indicate a correct classification of presence/absence of the ischemic lesion. The best values are bolded, and the second-best values are italicized.

	Team	BCT+ (n=13)		RICT+ (n=4)		R2CT+ (n=1)		BCT- (n=18) *		BCT-Con (n=3) *		RICT-Con (n=1) *	
		DICE↑	AVD↓	DICE↑	AVD↓	DICE↑	AVD↓	DICE↑	AVD↓	DICE↑	AVD↓	DICE↑	AVD↓
R1	<i>icomaia</i>	0.38 ± 0.25	42.66 ± 53.17	0.33 ± 0.27	29.62 ± 46.83	0.00	7.78	0.06 ± 0.17	30.76 ± 30.53	0.00 ± 0.00	14.01 ± 2.83	0.00	13.38
	<i>XMU</i>	<i>0.04 ± 0.10</i>	41.12 ± 50.15	0.02 ± 0.04	51.05 ± 56.41	0.00	0.60	0.00 ± 0.00	16.53 ± 28.35	1.00 ± 0.00	0.00 ± 0.00	1.00	0.00
	<i>LONGYING</i>	0.01 ± 0.01	43.95 ± 54.90	<i>0.03 ± 0.05</i>	<i>50.44 ± 55.79</i>	0.00	0.60	0.00 ± 0.00	15.68 ± 28.74	1.00 ± 0.00	0.00 ± 0.00	1.00	0.00
	<i>LaTIM</i>	0.00 ± 0.00	42.94 ± 56.35	0.00 ± 0.00	52.99 ± 58.89	0.00	0.60	0.00 ± 0.00	<i>16.06 ± 27.90</i>	1.00 ± 0.00	0.00 ± 0.00	1.00	0.00
R2	<i>icomaia</i>	0.38 ± 0.23	37.86 ± 47.38	0.32 ± 0.28	31.50 ± 55.53	0.00	7.18	0.07 ± 0.16	30.81 ± 29.64	0.00 ± 0.00	14.01 ± 2.83	0.00	6.09
	<i>XMU</i>	<i>0.04 ± 0.11</i>	<i>40.67 ± 45.12</i>	<i>0.02 ± 0.03</i>	50.97 ± 54.52	0.00	1.20	0.00 ± 0.00	17.00 ± 27.44	1.00 ± 0.00	0.00 ± 0.00	0.00	7.29
	<i>LONGYING</i>	0.01 ± 0.02	43.49 ± 49.43	<i>0.02 ± 0.04</i>	<i>50.36 ± 54.10</i>	0.00	1.20	0.00 ± 0.00	16.12 ± 27.84	1.00 ± 0.00	0.00 ± 0.00	0.00	7.29
	<i>LaTIM</i>	0.00 ± 0.00	42.63 ± 50.87	0.00 ± 0.00	52.91 ± 56.41	0.00	1.20	0.00 ± 0.00	<i>16.54 ± 26.84</i>	1.00 ± 0.00	0.00 ± 0.00	0.00	7.29

Table 3. Average DICE and AVD of the top four teams in NCCT segmentation based on NCCT group, as evaluated by two experts. Numbers in parentheses denote sample sizes for each group. There exists an overlapping of four cases in Control groups and the BCT- group, marked with an asterisk (*). The best values are bolded, and the second-best values are italicized.

Table 4 presents Dice scores and volume differences by lesion size, revealing *XMU*, *LONGYING*, and *LaTIM*'s high performance in the *no lesion* category. In contrast, *icomaia* predicted lesions inaccurately, with high volume differences for both raters. For *tiny* lesions, no team achieved overlap, with the lowest AVD reported by *LONGYING*. As for the *small* lesions, only *icomaia* estimated lesions overlapping with expert evaluations, nevertheless, other teams showed the lowest volume differences. *icomaia* also outperformed in *large* lesion evaluations with the highest Dice scores and significantly lower volume differences compared to other teams' averages around 90mL. Notably, non-zero Dice scores were observed for all teams, except for *LaTIM*.

Figure 4 summarizes the test set's average volume differences and Dice scores across all teams by radiologist annotations. Subjects, categorized and ordered by descending average lesion volume, showed the best performance in the *BCT+* and *RICT+* groups. Overlap was observed in nearly all cases, except for subjects with the smallest lesions in both groups. The *R2CT+* group saw no successful identification of small lesions, yet recorded lower volume differences. In *BCT-*, Dice scores dropped significantly due to the absence of clear visual indications of stroke lesions, but volume differences indicated less disparity. Control samples (*RICT-Con* and *BCT-Con*) showed good performance with a mean Dice score of 0.75 and volume differences of 3.50mL. Discrepancies in the presence of annotations led to varied results in the *RICT-Con* subject, with a notable Dice score difference for both radiologists (R1=0.75, R2=0) and volume differences (R1=3.35mL, R2=6.99mL). Interestingly, a single strategy predicted a lesion over this study, but such prediction did not correspond with the lesion delineated by R2.

Discussion and concluding remarks

The APIS challenge motivated participants to design and implement models to segment stroke lesions over NCCT studies, taking into consideration paired information between NCCT and ADC studies. At test, each model used only NCCT input observations and validated regarding the annotations from two radiologists. This work

	Team	NL (0mL, R1=4, R2=3)		T (≤1.19mL, R1=9, R2=6)		S (≤8.44mL, R1=7, R2=9)		M (≤42.38mL, R1=7, R2=9)		L (≥42.38mL, R1=9, R2=9)	
		DICE↑	AVD↓	DICE↑	AVD↓	DICE↑	AVD↓	DICE↑	AVD↓	DICE↑	AVD↓
R1	<i>icomaia</i>	0.00 ± 0.00	13.85 ± 2.33	0.00 ± 0.00	16.41 ± 18.68	0.17 ± 0.19	42.35 ± 65.02	0.28 ± 0.26	40.62 ± 51.05	0.45 ± 0.27	42.55 ± 28.47
	<i>XMU</i>	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	<i>0.57 ± 0.35</i>	0.00 ± 0.00	3.60 ± 1.48	<i>0.01 ± 0.02</i>	16.41 ± 5.55	<i>0.06 ± 0.12</i>	<i>91.73 ± 32.60</i>
	<i>LONGYING</i>	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.52 ± 0.36	0.00 ± 0.00	3.17 ± 1.79	0.00 ± 0.01	15.29 ± 8.66	0.02 ± 0.04	95.48 ± 36.65
	<i>LaTIM</i>	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	<i>0.57 ± 0.35</i>	0.00 ± 0.00	3.27 ± 1.89	0.00 ± 0.00	14.35 ± 8.52	0.00 ± 0.00	96.35 ± 38.59
R2	<i>icomaia</i>	0.00 ± 0.00	14.01 ± 2.83	0.00 ± 0.00	10.67 ± 5.13	0.08 ± 0.11	23.18 ± 35.44	0.27 ± 0.21	46.95 ± 54.61	0.44 ± 0.27	40.85 ± 33.87
	<i>XMU</i>	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	<i>0.78 ± 0.20</i>	0.00 ± 0.00	4.05 ± 2.06	<i>0.01 ± 0.02</i>	16.89 ± 9.22	<i>0.06 ± 0.14</i>	<i>87.34 ± 29.68</i>
	<i>LONGYING</i>	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.75 ± 0.22	0.00 ± 0.00	3.67 ± 1.85	0.00 ± 0.01	<i>15.97 ± 11.01</i>	0.02 ± 0.03	91.09 ± 32.64
	<i>LaTIM</i>	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	<i>0.78 ± 0.20</i>	0.00 ± 0.00	3.77 ± 2.18	0.00 ± 0.00	15.50 ± 11.11	0.00 ± 0.00	91.96 ± 33.99

Table 4. Average DICE and AVD of the top four teams in NCCT segmentation based on lesion size categories, as evaluated by two experts. Group names have been abbreviated as follows: NL for No Lesion, T for Tiny, S for Small, M for Medium, and L for Large. Values in parentheses indicate the volume range in mL and the number of samples in the group for each radiologist. The best values are bolded, and the second-best values are italicized.

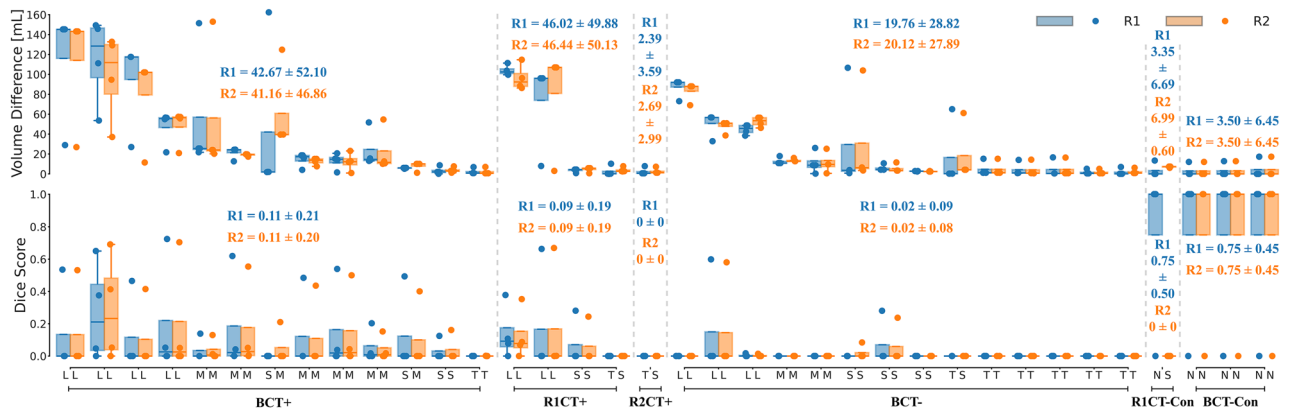


Fig. 4. Volume differences (top) and Dice scores (bottom) achieved by the participant strategies across test subjects, with respect to both expert annotations. The x-axis displays lesion categories based on each radiologist's mean lesion volume: N for no lesion, T for tiny, S for small, M for medium, and L for large.

reported and discussed the segmentation scores achieved by the four finalists, evidencing advances on lesion definition, but with remarkable challenges to discover and characterize lesions from NCCT. Other submitted solutions, employed Unet-like architectures, without the significant novelty of baseline methodology, so were excluded from the present study. Other submitted solutions, employed Unet-like architectures, without significant novelty of baseline methodology, so were excluded from the present study. Inspired by clinical procedure, we expect that learning strategies can learn prior information about stroke findings over different modalities, and then operate over simple NCCT studies. Hence, from such an open database is expected that the scientific community continue designing novel mechanisms to integrate stroke lesion findings on ADC to approximate lesion segmentations over NCCT studies. The APIS dataset targets small acute lesions within 24 hours of symptom onset, presenting a significant challenge that has not been extensively addressed in previous research or other datasets. In this study, it was conducted an analysis of computational tools regarding subgroups of samples, stratified according to the size of lesions. From this perspective, we hope to bring additional criteria to establish advantages and limitations of the proposed computational tools. In the literature, open databases and challenges have been dedicated to stroke segmentation providing NCCT studies with associated delineations²⁶, or with perfusion CT studies together with perfusion maps²⁷. Despite these efforts, the proposed strategies are focus on isolated modalities (AISD) with a low chance to carry out a technological transfer to operate in clinical domain. Also, other alternatives have approximated the variability of lesions but requires complementary perfusion maps, which is not the gold-standard during patient triage, diagnosis, and intervention. Besides, these studies lost the reported variability of expert annotations, which result critical to produce generalizable alternatives.

The analysis of agreement between radiologists revealed moderate consensus regarding the presence of stroke lesions on NCCT, evidenced by a kappa score of 0.72. This agreement was particularly notable in instances where both radiologists identified lesions (*BCT+*, 36.1%) or agreed on their absence (*BCT-*, 50%). In four instances (*R1CT+*), a lesion was identified by only one radiologist, while in one case (*R2CT+*), the other radiologist identified a lesion. For agreement on MRI delineations, the mean Dice score of 0.63 ± 0.29 and volume difference of 4.45 ± 7.24 mL underscored significant variations in expert delineations. Analysis of lesion volumes revealed a prevalence of smaller lesions, with R1 annotating slightly smaller volumes than R2. Annotation protocol emulates standard clinical practice, i.e., the radiologist uses clinical information in the admission records, which therefore may have introduced biases that influenced the delineation process. Specifically, detailed symptom descriptions might direct radiologists' attention to particular brain regions, potentially causing them to neglect other areas. Additionally, awareness of the duration between symptom onset and imaging may impact radiologists' interpretations of lesion characteristics, as they adjust their expectations based on the typical progression of ischemic damage over time. The *BCT+* group had median volumes indicative of *medium* sized lesions. In contrast, the *R1CT+* group included both *large* and *small* lesions. It is essential to point out that R2 overlooked these lesions in CT scans but successfully delineated them in ADC. In contrast, a single case in the *R2CT+* group had a *tiny* lesion, and the *BCT-* group predominantly contained *small* lesions. An interesting observation was made in the control cases (*BCT-Con*), where one case classified as control by one radiologist had a small lesion identified by the other (*R1CT-Con*). The scores in *BCT+* group showed the highest Dice score and a relatively low volume difference, while the *R2CT+* group achieved the lowest volume difference. These results illustrate the complexity and variability inherent in lesion segmentation, highlighting the clearer characterization of stroke lesions in ADC compared to NCCT, despite discrepancies in NCCT interpretations.

Further advancements are needed to effectively learn NCCT stroke lesion findings, which, according to conventional prognostic management of stroke patients, it should be the first imaging modality captured for stroke diagnosis⁴. The best scores for R1 and R2 reference annotations were comparable with results reported in the literature, but underscore the segmentation challenges inherent due to the poor contrast and hypoattenuation of ischemic lesions. The leaderboard reveals that the winning team outperformed other teams in traditional segmentation metrics, yet it demonstrated suboptimal performance in clinical metrics, primarily due to the algorithm's tendency to overestimate the quantity and volume of lesions. Furthermore, the overall rankings

indicate a performance gap between the leading team and other participants. The other participants, displayed comparable efficacy in both traditional segmentation and clinical metrics. Notably, XMU emerged as the superior performer in this subset, suggesting a more refined balance between segmentation precision and clinical utility.

The strategies employed by participants resulted in suboptimal overlap with expert annotations, as evidenced by uniformly low Dice scores. Disparities between precision and sensitivity in certain strategies imply a conservative prediction of lesion voxels, often leading to an absence of lesion annotations and a consequent increase in control patient classifications ($XMU = 27$, $LONGYING = 26$, $LaTIM = 29$). The high HD values observed across strategies indicate inadequate delineation of ischemic lesion boundaries, likely inflated by cases misclassified as controls, incurring the maximal euclidean distance penalty. Clinically, the results suggest a significant over-estimation of lesion count and size by the leading team (*icomaia*), with an ALD and AVD notably higher than other participants. This aspect raises concerns, as the volume of the ischemic core is recognized as a critical biomarker for determining a good diagnosis, treatment, and the prognosis of stroke patients^{57–59}. Conversely, other teams produced lesions more congruent with expert evaluations, revealing a divergence between traditional segmentation accuracy and clinical metrics, indicating a need for a multi-faceted evaluation. In the assessment of lesion-wise F1-scores, XMU achieved notable scores of 0.24 for R1 and 0.21 for R2. These results indicate a trend towards higher false positive rates in *icomaia*'s strategy, while other teams' methodologies tend to result in higher false negative rates of lesion segments. Regarding the computational strategy, all teams employed traditional autoencoder architectures for lesion segmentation. The leading team leveraged a transfer learning approach from the AISD dataset's public data, notable for its larger median lesion volume compared to the APIS training set. The substantial variation in training data likely skewed the winning team's algorithm towards producing larger lesion annotations. Consequently, the winning team outperformed other teams in standard segmentation and boundary delineation metrics but was less effective in clinically associated metrics. Interestingly, the *LaTIM* team integrated a generative module to produce synthetic ADC maps, operating as an intermediate step to achieve CT segmentations. This solution is promising and the low achieved results may be attributed to challenges in the learning process, the available set of data, and the associated difficulty of the lesion estimation after symptoms onset. The APIS challenge presents an opportunity for the scientific community to develop tailored mechanisms that emphasize cross-modality learning of lesion tissues, potentially enhancing stroke lesion segmentation from modalities with low sensitivity. Interestingly, the *LaTIM* team integrated a generative module to produce synthetic ADC maps, operating as an intermediate step to achieve CT segmentations. This solution is promising and the low achieved results may be attributed to challenges in the learning process, the available set of data, and the associated difficulty of the lesion estimation after symptoms onset. The APIS challenge presents an opportunity for the scientific community to develop tailored mechanisms that emphasize cross-modality learning of lesion tissues, potentially enhancing stroke lesion segmentation from modalities with low sensitivity.

Regarding the analysis of segmentation quality, instances where both radiologists concurred on lesion presence ($BCT+$), and when R1 indicated lesion presence ($RICT+$), exhibited superior segmentation performance. Conversely, suboptimal performance occurred where only R2 detected lesions ($R2CT+$) and in cases where both radiologists agreed on lesion absence ($BCT-$). The single $R2CT+$ instance, identified by an expert, resulted in inferior segmentation, likely influenced by the lesion's tiny size and the training strategies being adjusted solely on R1's annotations. This discrepancy underscores the significant impact of lesion visibility and expert consensus on the effectiveness of automated segmentation methods. Moreover, in scenarios where both radiologists did not annotate ischemic lesions ($BCT-Con$), XMU, LONGYING, and *LaTIM* teams obtained perfect metrics, contrasting with *icomaia*'s inaccurate lesion predictions. The $RICT-Con$ group, marked by radiologist disagreement, demonstrated varied performance across teams, reflecting the challenges posed by inconsistent expert annotations. Continuing the observed trend, *icomaia* consistently outperformed competitors in conventional segmentation metrics but evidenced limitations in clinical metrics. An exception to this pattern emerged in the $RICT+$ group, distinguished by the presence of larger lesions, where *icomaia*'s performance aligned more closely with clinical metric expectations. Importantly, it should be noted that the limited sample size in some NCCT groups may affect the reliability of the performance metrics. Future studies should include larger sample sizes to validate these findings more robustly.

According to lesion stratification (Table 4), the results indicate a correlation between increasing lesion size and improved overlap and higher volume difference metrics. These results highlight the challenge in accurately annotating smaller lesions and the increased discrepancies in volume prediction for larger lesions, suggesting the need for approaches tailored to different lesion scales in stroke segmentation tasks. Moreover, the trend continues with *icomaia* achieving higher Dice scores but larger volume differences. This suggests that the winning algorithm has a bias for marking larger lesions. The strategies of XMU and LONGYING began to show some alignment with lesions in the *medium* and *large* categories, respectively. *icomaia*'s strategy showed alignment with the *small* category annotations, achieving a Dice score of 0.17 ± 0.19 for R1 and 0.08 ± 0.11 for R2. The *tiny* category presented the most significant difficulties for all strategies. Figure 4 exhibited inherent challenges in devising segmentation strategies for IS lesions on NCCT scans. A detailed analysis reveals several instances where at least one strategy achieved significant overlap in cases with lesions identified on NCCT by experts. However, notable discrepancies were observed in volume estimates. Conversely, in only three cases where experts did not identify NCCT lesions, some overlap was observed, with one case showing overlap exclusively for R2. This indicates limitations in the strategies proposed for characterizing IS in instances where lesions were undetectable to experts. Similarly to the findings in Table 4, a direct correlation exists between lesion volume and volume difference, although this does not extend to the Dice score. This suggests that effective overlap is more dependent on the prominence of the visual patterns of ischemic lesions.

At the end of the APIS challenge, we requested participants to train models on ADC sequences and submit their solutions as an alternative solution. As expected, a remarked difference was observed, achieving on average Dice scores of 0.67 ± 0.29 and 0.58 ± 0.30 for the first and second radiologist, respectively. These results are

consistent with the literature²⁴, however, MRI is not as widely available in clinical practice as NCCT, being the first line allowing triage. Overall, the stroke segmentation over NCCT studies remains an open problem, which requires additional efforts to design new architectures and models to explore CT stroke lesion findings, taking advantage of complementary paired information. In particular, generative or foundational strategies could be integrated into traditional segmentation schemes to promote cross-modality information learning^{40,60,61}. We encourage the scientific community to contribute new ideas to improve the current solutions for stroke lesion segmentation on NCCT. The findings of this research emphasize the necessity of developing algorithms capable of precisely detecting small lesions. Additionally, the study's findings would be strengthened by the inclusion of additional data. Specifically, the small dataset size currently limits the generalizability of the results, highlighting the importance of a larger dataset for more reliable conclusions. Also, the significant performance improvements observed with transfer learning emphasize the need for a greater number and diversity of stroke samples. Future works should include more paired sequences, involving a comprehensive representation of stroke variability with medium to large stroke lesions. Additionally, diversifying the dataset with contributions from a broader range of institutions could significantly enrich the dataset, providing computational models access to a wider spectrum of stroke characteristics essential for training on paired sequences. In subsequent releases, clinical data will be made available alongside the images and lesion annotations, thereby enriching the analysis and aiding models in the segmentation of stroke lesions.

Data availability

The annotated dataset remains accessible to the public upon registration in the challenge: <https://bivl2ab.uis.edu.co/challenges/apis>. For access, contact the corresponding author, Fabio Martínez, at famarcar@saber.uis.edu.co.

Received: 21 February 2024; Accepted: 26 August 2024

Published online: 04 September 2024

References

1. Feigin, V. L. *et al.* World stroke organization (WSO): Global stroke fact sheet 2022. *Int. J. Stroke* **17**, 18–29 (2022).
2. Rennett, R. C. *et al.* Epidemiology, natural history, and clinical presentation of large vessel ischemic stroke. *Neurosurgery* **85**, S4–S8. <https://doi.org/10.1093/neuros/nyz042> (2019).
3. Powers, W. J. *et al.* Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the american heart association/american stroke association. *Stroke* **50**, e344–e418 (2019).
4. Harpaz, D., Eltzov, E., Seet, R. C., Marks, R. S. & Tok, A. I. Point-of-care-testing in acute stroke management: An unmet need ripe for technological harvest. *Biosensors* **7**, 30 (2017).
5. Potter, C. A. *et al.* Ct for treatment selection in acute ischemic stroke: A code stroke primer. *Radiographics* **39**, 1717–1738 (2019).
6. Wardlaw, J. M. *et al.* Accuracy of automated computer-aided diagnosis for stroke imaging: A critical evaluation of current evidence. *Stroke* **53**, 2393–2403. <https://doi.org/10.1161/STROKEAHA.121.036204> (2022).
7. Sarraj, A. *et al.* Trial of endovascular thrombectomy for large ischemic strokes. *New Engl. J. Med.* **388**, 1259–1271. <https://doi.org/10.1056/NEJMoa2214403> (2023).
8. Rubin, M. N. & Barrett, K. M. What to do with wake-up stroke. *Neurohospitalist* **5**, 161–172 (2015).
9. Zhang, J. *et al.* Use of dwi-flair mismatch to estimate the onset time in wake-up strokes. *Neuropsychiatr. Dis. Treat.* **18**, 355 (2022).
10. Bevers, M. B. *et al.* Apparent diffusion coefficient signal intensity ratio predicts the effect of revascularization on ischemic cerebral edema. *Cerebrovasc. Dis.* **45**, 93–100 (2018).
11. Martel, A. L. *et al.* Measurement of infarct volume in stroke patients using adaptive segmentation of diffusion weighted MR images. *Lect. Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*[SPACE]https://doi.org/10.1007/10704282_3 (1999).
12. Rana, A. K., Wardlaw, J. M., Armitage, P. A. & Bastin, M. E. Apparent diffusion coefficient (ADC) measurements may be more reliable and reproducible than lesion volume on diffusion-weighted images from patients with acute ischaemic stroke-implications for study design. *Magnet. Resonan. Imaging* **21**, 617–624 (2003).
13. Mikhail, P., Le, M. G. D. & Mair, G. Computational image analysis of nonenhanced computed tomography for acute ischaemic stroke: A systematic review. *J. Stroke Cerebrovasc. Dis.* **29**, 104715. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.104715> (2020).
14. Murray, N. M., Unberath, M., Hager, G. D. & Hui, F. K. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review. *J. NeuroIntervent. Surg.* **12**, 156–164. <https://doi.org/10.1136/neurintsurg-2019-015135> (2020).
15. Brinjikji, W. *et al.* e-ASPECTS software improves interobserver agreement and accuracy of interpretation of aspects score. *Intervent. Neuroradiol.* **27**, 781–787. <https://doi.org/10.1177/15910199211011861> (2021).
16. Chan, N. *et al.* Machine-learning algorithm in acute stroke: Real-world experience. *Clin. Radiol.* **78**, e45–e51 (2023).
17. Adhya, J. *et al.* Positive predictive value and stroke workflow outcomes using automated vessel density (rapid-cta) in stroke patients: One year experience. *Neuroradiol. J.* **34**, 476–481 (2021).
18. Al-Kawaz, M., Primiani, C., Urrutia, V. & Hui, F. Impact of rapidai mobile application on treatment times in patients with large vessel occlusion. *J. NeuroIntervent. Surg.* **14**, 233–236 (2022).
19. Dekeyser, S. *et al.* Distinction between contrast staining and hemorrhage after endovascular stroke treatment: One CT is not enough. *J. NeuroIntervent. Surg.* **9**, 394–398 (2017).
20. Liew, S.-L. *et al.* A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Sci. Data* **5**, 1–11 (2018).
21. Liew, S.-L. *et al.* A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Sci. Data* **9**, 320 (2022).
22. Maier, O. *et al.* Isles 2015—A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med. Image Anal.* **35**, 250–269 (2017).
23. Winzeck, S. *et al.* Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. *Front. Neurol.* **9**, 679 (2018).
24. Hernandez Petzsche, M. R. *et al.* Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Sci. Data* **9**, 762 (2022).
25. Liu, C.-F. *et al.* A large public dataset of annotated clinical MRIS and metadata of patients with acute stroke. *Sci. Data* **10**, 548 (2023).

26. Liang, K. et al. Symmetry-enhanced attention network for acute ischemic infarct segmentation with non-contrast CT images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. 432–441 (Springer, 2021).
27. Hakim, A. et al. Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: Lessons from the isles challenge. *Stroke* **52**, 2328–2337 (2021).
28. Lin, L., Bivard, A., Levi, C. R. & Parsons, M. W. Comparison of computed tomographic and magnetic resonance perfusion measurements in acute ischemic stroke: Back-to-back quantitative analysis. *Stroke* **45**, 1727–1732 (2014).
29. Lansberg, M. G. et al. MRI profile and response to endovascular reperfusion after stroke (defuse 2): A prospective cohort study. *Lancet Neurol.* **11**, 860–867 (2012).
30. Allmendinger, A. M., Tang, E. R., Lui, Y. W. & Spektor, V. Imaging of stroke: Part 1, perfusion CT overview of imaging technique, interpretation pearls, and common pitfalls. *Am. J. Roentgenol.* **198**, 52–62 (2012).
31. Kuang, H., Menon, B. K. & Qiu, W. Automated infarct segmentation from follow-up non-contrast CT scans in patients with acute ischemic stroke using dense multi-path contextual generative adversarial network. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. 856–863 (Springer, 2019).
32. Wang, Y., Katsaggelos, A. K., Wang, X. & Parrish, T. B. A deep symmetry convnet for stroke lesion segmentation. In *2016 IEEE International Conference on Image Processing (ICIP)*. 111–115 (IEEE, 2016).
33. Ni, H. et al. Asymmetry disentanglement network for interpretable acute ischemic stroke infarct segmentation in non-contrast CT scans. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022* (Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S. eds.). Vol. 13438. 416–426 https://doi.org/10.1007/978-3-031-16452-1_40 (Springer Nature Switzerland, 2022) (series title: Lecture Notes in Computer Science).
34. Xu, Z. & Ding, C. Combining convolutional attention mechanism and residual deformable Transformer for infarct segmentation from CT scans of acute ischemic stroke patients. *Front. Neurol.* **14**, 1178637. <https://doi.org/10.3389/fneur.2023.1178637> (2023).
35. Tureckova, A. & Rodríguez-Sánchez, A. J. ISLES challenge: U-shaped convolution neural network with dilated convolution for 3D stroke lesion segmentation. *Lect. Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11383 LNCS**, 319–327 https://doi.org/10.1007/978-3-030-11723-8_32 (2019).
36. Dolz, J., Ben Ayed, I. & Desrosiers, C. Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities. *Lect. Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11383 LNCS**, 271–282 https://doi.org/10.1007/978-3-030-11723-8_27. arXiv:1810.07003 (2019).
37. Clérigues, A. et al. Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. *Comput. Biol. Med.* **115**, 103487. <https://doi.org/10.1016/j.compbiomed.2019.103487> (2019).
38. Liu, P. Stroke lesion segmentation with 2D novel CNN pipeline and novel loss function. *Lect. Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11383 LNCS**, 253–262 https://doi.org/10.1007/978-3-030-11723-8_25 (2019).
39. Wang, G. et al. Automatic ischemic stroke lesion segmentation from computed tomography perfusion images by image synthesis and attention-based deep neural networks. *Med. Image Anal.* **65**, 101787 <https://doi.org/10.1016/j.media.2020.101787>. arXiv:2007.03294 (2020).
40. Gutierrez, A. et al. Lesion-preserving unpaired image-to-image translation between MRI and CT from ischemic stroke patients. *Int. J. CARS* **18**, 827–836. <https://doi.org/10.1007/s11548-022-02828-4> (2023).
41. Kuang, H., Menon, B. K., Sohn, S. I. & Qiu, W. EIS-Net: Segmenting early infarct and scoring ASPECTS simultaneously on non-contrast CT of patients with acute ischemic stroke. *Med. Image Anal.* **70**, 101984. <https://doi.org/10.1016/j.media.2021.101984> (2021).
42. Gauriau, R. et al. Head CT deep learning model is highly accurate for early infarct estimation. *Sci. Rep.* **13**, 189 (2023).
43. Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: Dicom to Nifti conversion. *J. Neurosci. Methods* **264**, 47–56 (2016).
44. Lutkenhoff, E. S. et al. Optimized brain extraction for pathological brains (optibet). *PLoS one* **9**, e115551 (2014).
45. Shamonin, D. P. et al. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer’s disease. *Front. Neuroinform.* **7**, 50 (2014).
46. Maier-Hein, L. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 5217 (2018).
47. Iglesias, J. E. et al. Synthsr: A public AI tool to turn heterogeneous clinical brain scans into high-resolution t1-weighted images for 3D morphometry. *Sci. Adv.* **9**, eadd3607 (2023).
48. Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. & Hoffmann, M. Synthstrip: Skull-stripping for any brain image. *NeuroImage* **260**, 119474 (2022).
49. Rorden, C., Bonilha, L., Fridriksson, J., Bender, B. & Karnath, H.-O. Age-specific CT and MRI templates for spatial normalization. *Neuroimage* **61**, 957–965 (2012).
50. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. NNU-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
51. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. 6105–6114 (PMLR, 2019).
52. Dai, Z., Liu, H., Le, Q. V. & Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **34**, 3965–3977 (2021).
53. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778 (2016).
54. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988 (2017).
55. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134 (2017).
56. Hatamizadeh, A. et al. Unetr: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 574–584 (2022).
57. Laredo, C. et al. Prognostic significance of infarct size and location: The case of insular stroke. *Sci. Rep.* **8**, 9498 (2018).
58. Kaczmarczyk, K., Wit, A., Krawczyk, M., Zaborski, J. & Gajewski, J. Associations between gait patterns, brain lesion factors and functional recovery in stroke patients. *Gait Posture* **35**, 214–217 (2012).
59. Sperber, C., Gallucci, L., Mirman, D., Arnold, M. & Umarova, R. M. Stroke lesion size—Still a useful biomarker for stroke severity and outcome in times of high-dimensional models. *NeuroImage Clin.* **40**, 103511 (2023).
60. Abu-Srhan, A., Almallahi, I., Abushariah, M. A., Mahafza, W. & Al-Kadi, O. S. Paired-unpaired unsupervised attention guided GAN with transfer learning for bidirectional brain MR-CT synthesis. *Comput. Biol. Med.* **136**, 104763 (2021).
61. Wang, B. et al. Quantitative cerebral blood volume image synthesis from standard MRI using image-to-image translation for brain tumors. *Radiology* **308**, e222471. <https://doi.org/10.1148/radiol.222471> (2023).

Acknowledgements

The authors thank Ministry of science, technology and innovation of Colombia (MINCIENCIAS) for supporting this research work by the project “Mecanismos computacionales de aprendizaje profundo para soportar tareas de localización, segmentación y pronóstico de lesiones asociadas con accidentes cerebrovasculares isquémicos”, with code 91934.

Author contributions

Conceptualization: S.G., E.R., D.M., F.M.; Challenge Methodology: S.G., E.R., F.M.; Evaluation platform: S.G., E.R.; Data acquisition: S.G., E.R., D.M., A.O., F.M.; Data curation: S.G., E.R., D.M., A.O., F.M.; Data analysis: S.G., D.M., A.O., F.M.; Deep learning methods: E.D.L.R., J.S., J.K., Y.L., M.E.H.D.; Validation: S.G., F.M.; Visualization and figures: S.G., E.R.; Writing, coordination/editing: S.G., F.M.; Writing, results: S.G., F.M.; Writing, deep learning methods: S.G., E.D.L.R., J.S., J.K., Y.L., M.E.H.D.; Writing, discussion: S.G., F.M.; Project administration: D.M., P.C., F.M.; Funding acquisition: D.M., P.C., F.M. All authors have reviewed and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-71273-x>.

Correspondence and requests for materials should be addressed to F.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024