BONBID-HIE 2023: Lesion Segmentation Challenge in BOston Neonatal Brain Injury Data for Hypoxic Ischemic Encephalopathy

Rina Bao, Anna N. Foster, Ya'nan Song, Rutvi Vyas, Ankush Kesri, Imad Eddine Toubal, Elham Soltani Kazemi, Gani Rahmon, Taci Kucukpinar, Mohamed Almansour, Mai-Lan Ho, K. Palaniappan, Dean Ninalga, Chiranjeewee Prasad Koirala, Sovesh Mohapatra, Gottfried Schlaug, Marek Wodzinski, Henning Muller, David G. Ellis, Michele R. Aizenberg, M. Arda Aydın, Elvin Abdinli, Gozde Unal, Nazanin Tahmasebi, Kumaradevan Punithakumar, Tian Song, Yun Peng, Sara V. Bates, Randy Hirschtick, P. Ellen Grant, and Yangming Ou

***Abstract*— Hypoxic Ischemic Encephalopathy (HIE) represents a brain dysfunction, affecting approximately 1 to 5 per 1000 full-term neonates. The precise delineation and segmentation of HIE-related lesions in neonatal brain Magnetic Resonance Images (MRI) are pivotal in advancing outcome predictions, identifying patients at high risk, elucidating neurological manifestations, and assessing treatment efficacies. Despite its importance, the development of algorithms for segmenting HIE lesions from MRI volumes has been impeded by data scarcity. Addressing this critical gap, we organized the first BONBID-HIE challenge with diffusion MRI data (Apparent Diffusion Coefficient (ADC) maps) for HIE lesion segmentation, in conjunction with the MICCAI 2023. Totally 14 algorithms were submitted, employing a gamut of cutting-edge automatic machine-learning-based segmentation algorithms. Our comprehensive analysis of HIE lesion segmentation and submitted algorithms facilitates an in-depth evaluation of the current technological zenith, outlines directions for future advancements, and highlights persistent hurdles. To foster ongoing research and benchmarking, the annotated HIE dataset, developed algorithm dockers, and unified evaluation codes are accessible through a dedicated online platform (https://bonbid-hie2023.grand-challenge.org).**

This work was funded, in part, by the Harvard Medical School and Boston Children's Hospital through Early Career Development Fellowship, Thrasher Research Fund Early Career Awards, NIH R21NS121735, R61NS126792, and R03HD104891.

Rina Bao and Yangming Ou are with Boston Children's Hospital and Harvard Medical School, USA (e-mails: rina.bao@childrens.harvard.edu; yangming.ou@childrens.harvard.edu). Corresponding authors: Rina Bao and Yangming Ou. Anna N. Foster, Ya'nan Song, Rutvi Vyas, and Ankush Kesri are with Boston Children's Hospital, USA. Sara V. Bates, Randy Hirschtick, and P. Ellen Grant are with Harvard Medical School and Massachusetts General Hospital, USA. Imad Eddine Toubal, Elham Soltani Kazemi, Gani Rahmon, Taci Kucukpinar, Mohamed Almansour, Mai-Lan Ho, and K. Palaniappan are with the University of Missouri-Columbia, USA. Dean Ninalga is with the University of Toronto, Canada. Chiranjeewee Prasad Koirala, Sovesh Mohapatra, and Gottfried Schlaug are with the University of Massachusetts Amherst, USA. Marek Wodzinski is with the University of Applied Sciences Western Switzerland, Switzerland, and AGH University of Krakow, Poland. Henning Muller is with the University of Geneva, Switzerland. David G. Ellis and Michele R. Aizenberg are with the University of Nebraska Medical Center, USA. M. Arda Aydın and Gozde Unal are with Istanbul Technical University, Turkey. Elvin Abdinli is with the Technical University of Munich, Germany. Nazanin Tahmasebi and Kumaradevan Punithakumar are with the University of Alberta, Canada. Tian Song is with C&TS Philips, China. Yun Peng is with The Second Affiliated Hospital of Nanchang University and Cihuai Cardio Cerebrovascular Hospital, China.

Fig. 1. There are 180 ongoing trials related to HIE spreading over 33 countries and 5 continents (data retrieved on clinicaltrials.gov on 11/2024).

***Index Terms*— Brain injury, Lesion segmentation, Machine Learning, MRI, Challenge, Benchmark, Hypoxic Ischemic Encephalopathy, Algorithm Comparison, Algorithm development**

I. Introduction

NEONATAL hypoxic-ischemic encephalopathy (HIE) remains a significant public health concern, characterized by brain injury due to insufficient blood and oxygen supply to the brain. HIE affects approximately 1 to 5 per 1000 term-born neonates worldwide each year, with an estimated annual cost exceeding $2 billion in the United States alone, not accounting for the substantial burden on affected families [1]–[3]. Despite the adoption of Therapeutic Hypothermia (TH) as the standard of care, there is a substantial proportion of HIE patients (35%–50%) experiencing adverse neurocognitive outcomes [4]–[6]. Reducing mortality and morbidity associated with HIE is, therefore, a critical public health objective. Globally, there are currently 180 ongoing clinical trials related to HIE, spanning 33 countries and five continents (Figure 1, data retrieved on clinicaltrials.gov on 11/2024) [7]–[13]. Early identification of patients at high risk for adverse outcomes sooner after initiation of therapy remains a significant challenge, as clinical outcomes are often not reliably measurable until the age of two years [14]. This highlights the urgent

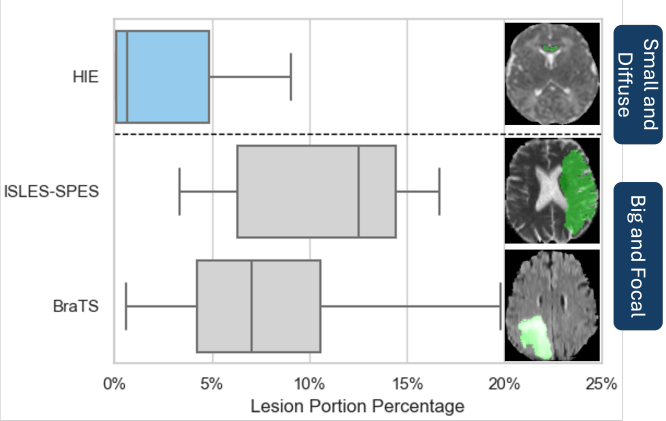

Fig. 2. Comparison of small, diffuse lesions in HIE [20] with large, focal lesions in ISLES-SPES (acute stroke outcome/penumbra estimation) [21] and BraTS [22]. The boxplot quantifies the percentage of brain volume affected by lesions across all patients in each dataset.

need for accurate and reliable biomarkers to enable an early prognosis and outcome prediction.

Accurate identification and segmentation of HIE-related lesions in neonatal brain magnetic resonance images (MRIs) is a critical step toward this goal. Clinical trials often rely on the NRN scoring system [15], [16], which is based on expert assessments of lesion extent and location, to predict neurocognitive outcomes. Lesion patterns in regions such as the basal ganglia, thalamus, and watershed areas are associated with distinct neurocognitive impairments, including motor, language, and executive dysfunction [16]–[19].

Machine learning, especially deep learning, in the detection and segmentation of lesions remains largely unexplored in HIE. Developing robust segmentation algorithms is challenging due to two interrelated obstacles:*(i) Data scarcity and annotation limitations.* Compared to neurological disorders such as brain tumors [23], [24], Alzheimer's disease [25], [26], and ischemic stroke [21], [27], [28], neonatal HIE lacks sufficient publicly available imaging benchmarks. High-quality datasets containing annotated MRIs alongside clinical and outcome information remain rare. One reason is that integrating imaging with longitudinal clinical and neurodevelopmental outcomes requires long-term follow-up, which is often challenging in both clinical and research settings. Additionally, expert annotation of HIE lesions demands specialized knowledge of neonatal neuroimaging, further limiting the scalability of dataset curation. *(ii) Algorithmic challenges posed by lesion characteristics.* Unlike tumors [22], HIE lesions are typically small ($< 1\%$) and diffuse (multi-focal) [20], [29], with over half of our patient cohort exhibiting such characteristics. This makes segmentation of HIE MRI data more challenging compared to tasks involving adult brain tumors, where lesions are generally larger and more focal, as illustrated in Figure 2. These characteristics lead to extreme class imbalance and an increased risk of false negatives in conventional segmentation pipelines.

To tackle these challenges, we collected a cohort of 133 cases with expert-annotated lesions over a decade of dedicated research [1], [20], forming the the Boston Neonatal Brain Injury Dataset for Hypoxic-Ischemic Encephalopathy (BONBID-HIE) [20]. To accelerate advancements in this domain, we organized the BONBID-HIE Lesion Segmentation Challenge. This challenge provided a direct, fair and independently controlled comparison of automated methods on this rigorously curated public dataset and platform. The challenge was conducted as a satellite event at the 26th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2023) in Vancouver, Canada. The event garnered global interest, with more than 140 registrations and 14 successful submissions in the test phase, including docker containers for algorithms and comprehensive method descriptions, which were published in our workshop proceedings [30].

In summary, this paper introduces the BONBID-HIE Lesion Segmentation Challenge. Despite the limited availability of data, the ultimate goal was to alleviate the burden on medical experts in diagnosing and prognosing HIE. The challenge includes the publicly accessible BONBID-HIE dataset, the submitted algorithm containers and their corresponding results, and the accompanying online validation tools as ongoing benchmarking resources. Our primary goal was to provide brain MRIs for HIE to inspire new research directions in HIE lesion segmentation. The event provided a unique platform for participants to explore machine learning methods and their practical applications in segmenting small and diffuse lesions in HIE. Concurrently, the challenge aimed to foster interdisciplinary research and collaboration, further advancing efforts in HIE outcome prediction. Additionally, we hope that this dataset will also be valuable for other small and diffuse lesion segmentation tasks.

The paper is organized as follows: Section 2 describes the BONBID-HIE challenge, including data characteristics, manual lesion annotations, and the evaluation process. Section 3 presents the challenge results, along with a statistical analysis of the performance of different algorithms. Section 4 discusses failure cases, method limitations, and future directions for HIE challenges. Section 5 concludes the paper.

II. 1st BONBID-HIE Lesion Segmentation Challenge

The 1st BONBID-HIE Lesion Segmentation Challenge was held as an online challenge and workshop in conjunction with MICCAI 2023. It was designed to promote continuous engagement by allowing new groups to access the training and test data, submit their segmentations, and automatically compare and rank their results against all previous submissions.

A. Data Settings and Annotations

Setting. 133 cases were retrospectively collected from a cohort of neonates diagnosed with HIE at Massachusetts General Hospital with Institutional Review Board approval from Massachusetts General Hospital and Boston Children's Hospital to anonymize, curate, annotate, and release them [20]. The image data was divided into three sets: 85 cases for training, 4 cases for docker sanity validation, and 44 cases for

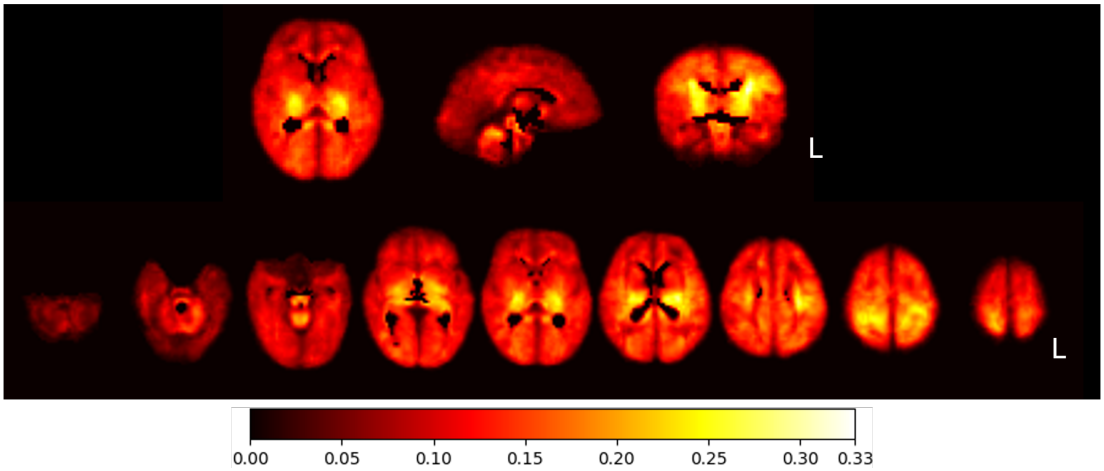

Fig. 3. Statistical lesion atlas quantifying the voxel-wise lesion frequency in our cohort of N=133 patients in the normal 0-14 days ADC atlas space from BONBID-HIE dataset [20].

testing. For each case, the provided inputs included Apparent Diffusion Coefficient (ADC) maps and Z_{ADC}, which were used as inputs for the algorithm containers. The output of the algorithm was the corresponding binary lesion segmentation map for each case.

Image Preprocessing. MRIs were acquired on either GE 1.5T Signa scanner (N=52, scanned during 2001-2012) or SIEMENS 3T Trio scanner (N=81, scanned during 2012-2018). Diffusion tensor imaging has the protocol as follows: TR=$7500-9500$ ms, TE=$80-115$ ms, b=$1000 s/mm^2$, voxel size=$2 \times 2 \times 2\ mm^3$ and 30 diffusion directions (SIEMENS scanner) or $1.5 \times 1.5 \times (4.0 - 6.0)\ mm^3$ and 6 diffusion directions (GE scanner). The MRI preprocessing pipeline applied in this study was previously established and validated for the BONBID-HIE dataset [20]. Specifically, preprocessing included N4 bias field correction to address intensity nonuniformities [31], field-of-view normalization to ensure consistent spatial coverage across images [32], and multi-atlas skull stripping tailored explicitly for ADC maps [33]. For additional details, please refer to [20]. All images retained their original voxel sizes without resampling. Consequently, segmentation masks were evaluated in each case's voxel space. This preserved resolution, avoided interpolation artifacts, and maintained clinically relevant spatial accuracy.

ADC Maps. Diffusion-weighted imaging (DWI) is widely used in brain MRI to probe the microstructural properties of tissue by sensitizing the MRI signal to water molecule motion [34]. The ADC is a quantitative metric derived from DWI that reflects the magnitude of water diffusion within a voxel. ADC is computed from DWI acquired with different b-values by modeling the signal attenuation according to a mono-exponential relationship: $S(b) = S_0 \cdot e^{-b \cdot \mathrm{ADC}}$, where $S(b)$ and S_0 denote the signal intensities at b-value b and at $b = 0$, respectively. Taking the natural logarithm of this equation and fitting a linear regression across the acquired b-values yields the ADC estimate: $\mathrm{ADC} = -\frac{\ln(S(b)/S_0)}{b}$. DWI and the resulting ADC maps are an appropriate modality to detect hypoxic ischemic brain injury in neonates diagnosed with HIE in the first few days after therapeutic hypothermia. Hypoxic ischemic injury leads to restricted water diffusion, which appears bright on DWI and dark on ADC maps, with these signal changes occurring earlier and being more clearly visible than signal changes on conventional T1- or T2-weighted sequences [35]–[38]. Therefore, DWI and ADC maps offer higher early sensitivity in detecting HIE related brain injury [39]–[42].

Z_{ADC} Maps. We developed Z_{ADC} maps to normalize and make ADC values comparable across brain voxel locations [20]. Z_{ADC} maps quantify location-specific deviations from normal, which is important for abnormal region segmentation. To generate Z_{ADC} maps, the following steps are performed: (1) A normative ADC atlas is constructed from the scans of 13 neonates, capturing the mean and standard deviation of ADC values at each voxel [43]. (2) A deformation field D is computed, which maps each voxel x in the patient's ADC map to its anatomically corresponding location $D(x)$ in the atlas space. The normal range of ADC variation is defined by the mean $\mu D(x)$ and standard deviation $\sigma D(x)$ for each voxel. (3) The patient's ADC value I_x at voxel x is converted to a Z-score:

$$Z_{ADC}(x) = \frac{I_x - \mu_D(x)}{\sigma_D(x)}. \tag{1}$$

Therefore, the Z_{ADC} value at voxel location x quantifies how many standard deviations the patient's ADC value I_x deviates from the mean normal ADC value at that anatomical location. For details on this process, please refer to [20].

Annotations. HIE lesions were initially identified in radiology reports using ADC maps and structural MRIs as primary imaging sequences [44], [45]. To translate these descriptions into voxel-level annotations, a two-step expert consensus approach was employed. First, lesions were manually annotated on ADC maps based on neuroradiology reports, with adjustments for integrity across axial, coronal, and sagittal planes. In cases of uncertainty or disagreement, consensus was reached through discussions among three experienced pediatric neuroradiologists. This multi-expert consensus process provided a robust and unbiased set of lesion annotations, marking the first ADC-based HIE lesion annotations [20].

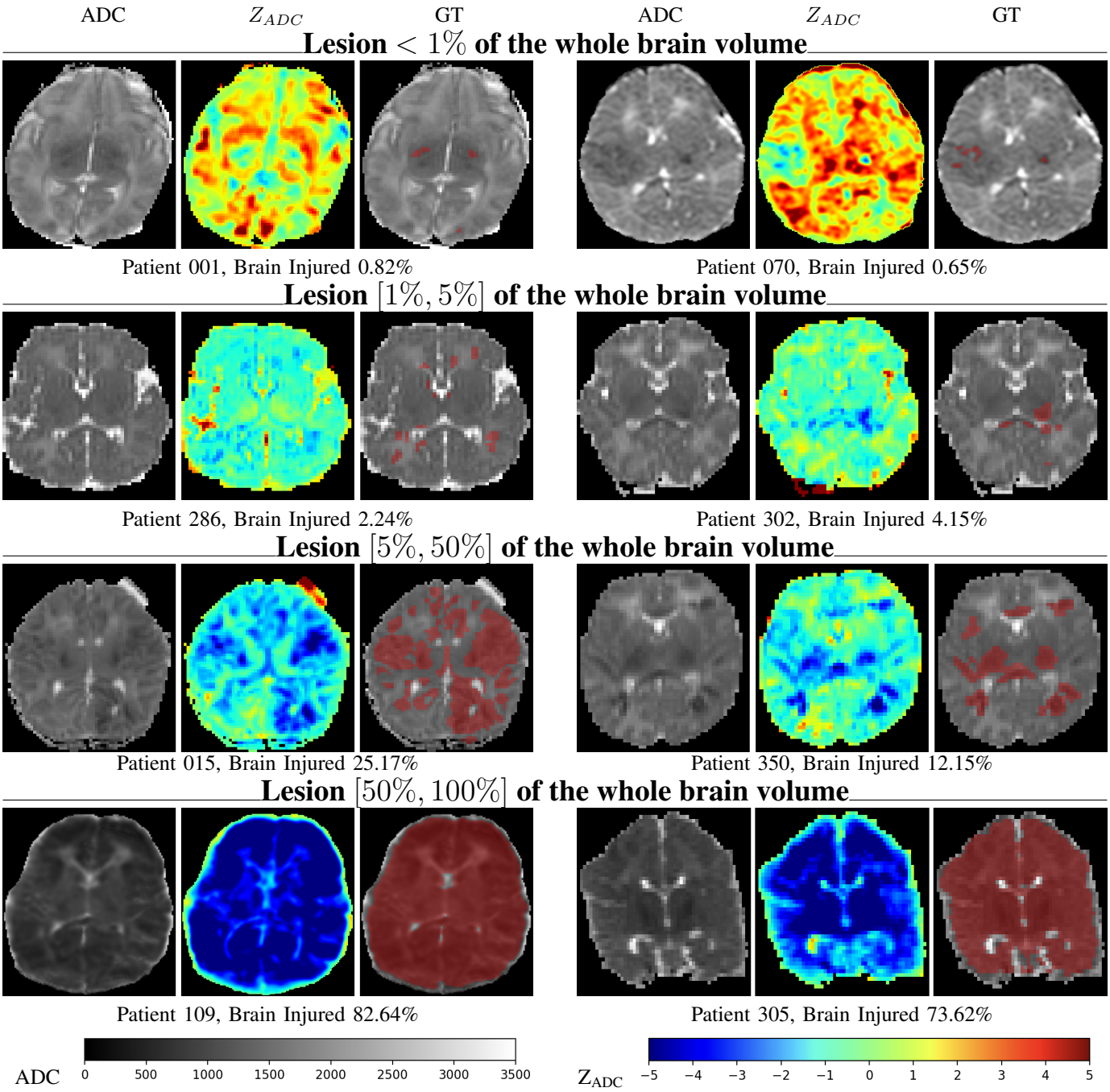

Fig. 4. Representative cases from the BONBID-HIE dataset visualized across a spectrum of brain injury percentages. Each row shows axial slices from two patients with similar lesion percentage categories. For each patient, three images are shown: the original ADC map, the corresponding Z_{ADC} map, and the expert-annotated ground truth (GT) lesion mask. Lesions are categorized based on their volume relative to the whole brain: $< 1\%$, [1%, 5%), [5%, 50%), and [50%, 100%). The percent brain injury is indicated below each case. This figure illustrates the wide variability in lesion extent and appearance, highlighting the challenge of segmenting small and multifocal HIE lesions.

B. Dataset Characteristics

Lesion Statistics. Figure 3 illustrates the spatial distribution of HIE lesions in neonates, presented as a frequency map overlaid on a normal brain atlas. The overlay on the normal atlas provides a clear visual comparison, illustrating the common regions impacted by HIE and their deviation from typical brain ADC values. As shown in the figure, HIE lesions are predominantly concentrated in the thalamus, perirolandic cortex, expected location of the corticospinal tract, midbrain and superior vermis.

Lesion Distribution. All Lesion percentages in our paper refer to the proportion of predicted lesion volume relative to the whole brain volume. As shown in Table I, more than 50% of HIE lesions are $< 1\%$ of the brain. This imbalance challenges machine learning algorithms, limiting performance optimization. In Figure 4, we illustrate the varying percentages of brain injury in HIE, displaying corresponding ADC, Z_{ADC}, and lesion segmentation maps.

TABLE I
DISTRIBUTION OF BRAIN INJURY EXTENT IN NEONATES WITH HIE

Lesion percentage	# of cases	Percentage
<1%	74	55.64%
[1%, 5%)	26	19.55%
[5%, 100%)	33	5.26%

C. Training and Testing

TABLE II
NUMBER OF PATIENTS SCANNED BY DIFFERENT SCANNERS IN PATIENT SUBGROUPS WITH DIFFERENT LESION VOLUMES.

Split	Scanner	% of brain volume lesioned $< 1\%$	$[1\%, 5\%]$	$> 5\%$	Total
Train&Val	GE 1.5T	15	2	16	33
	SIEMENS 3T	37	12	7	56
Test	GE 1.5T	8	5	6	19
	SIEMENS 3T	14	7	4	25

The postmenstrual age (PMA) at the time of MRI scan was 4.2 ± 2.6 days for the training set and 3.3 ± 1.3 days for the testing set. Due to the limited number of cases, we prioritized balancing lesion percentage and scanner types across the training and testing subsets. Lesion distribution was stratified as follows: in the training set, 46 small lesions ($< 1\%$), 16 medium lesions (1–5%), and 25 large lesions ($> 5\%$); and in the testing set, 22 small, 12 medium, and 10 large lesions (Table II). The training and testing data were also split while taking into account two scanner types (SIEMENS 3T and GE 1.5T) to ensure variability and robustness under different acquisition conditions.

D. Data and Evaluation Availability

One of the primary objectives of the BONBID-HIE is to provide an open-source repository to support the continuous development of algorithms. The BONBID-HIE lesion segmentation dataset is publicly available through Zenodo [1]. A unified portal for accessing benchmarks, submitting algorithm Docker containers, and performing automatic evaluations is available at challenge website [2]. All submitted algorithm dockers can also be accessed here. The evaluation codes are available on the challenge's GitHub repository [3].

E. Thresholding Z_{ADC} as baseline method

We provide a baseline method for the BONBID-HIE 2023 challenge. This method was originally described in our dataset paper [20]. Z_{ADC} quantifies how many standard deviations the patient's ADC value (I_x) deviates from the mean normal ADC value at the corresponding anatomical location. As reported in [20], by simply thresholding Z_{ADC} at -2, the DICE for lesion prediction using this method achieves a value of 0.54 on the entire dataset (N=133). When evaluated on this test set (N=44), the DICE score of Z_{ADC} (-2) is 0.58 (Table IV).

[1]https://zenodo.org/records/10602767
[2]https://bonbid-hie2023.grand-challenge.org/
[3]https://github.com/baorina/BONBID-HIE-MICCAI2023/tree/main

F. Evaluation metrics and ranking

Evaluation metrics. Evaluation metrics are DICE coefficient (DICE), and Mean Average Surface Distance (MASD) [53] and Normalized Surface Distance (NSD) [53] in different percentages of lesion ($< 1\%$, $1\% \sim 5\%$, $> 5\%$).

DICE. The DICE coefficient evaluates the overlap between two segmented regions, measuring the similarity between the predicted and ground truth segments, with a higher value indicating better agreement. DICE is computed as,

$$DICE = \frac{2 \times |P \cap Q|}{|P| + |Q|}, \tag{2}$$

where $|P|$ and $|Q|$ are the predicted segmentation lesion volume and ground truth lesion volume, respectively.

MASD. MASD measures the average distance between the surfaces of the predicted and ground truth segmentations, assessing how closely the boundaries align, with a lower value indicating better performance. The MASD between two sets of boundary points S_P of P and S_Q of Q is calculated as:

$$\text{MASD}(S_P, S_Q) = \frac{1}{2}\left(\frac{\sum_{p \in S_P} d(p, S_Q)}{|S_P|} + \frac{\sum_{q \in S_Q} d(q, S_P)}{|S_Q|}\right) \tag{3}$$

where: $d(p, S_Q) = \min_{q \in S_Q} d(p, q)$ and $d(q, S_P) = \min_{p \in S_P} d(q, p)$. $d(p, q)$ is the Euclidean distance between points p and q. $|S_P|$ and $|S_Q|$ denote the number of points in sets S_P and S_Q, respectively. $\frac{\sum_{p \in S_P} d(p, S_Q)}{|S_P|}$ is the average of the minimum distances from each point in S_P to the set S_Q. $\frac{\sum_{q \in S_Q} d(q, S_P)}{|S_Q|}$ is the average of the minimum distances from each point in S_Q to the set S_P. Therefore, the MASD is the mean of these two average surface distances, providing a symmetric measure of the average surface distance between the two boundaries.

NSD. NSD evaluates the proportion of surface points on the predicted segmentation that lie within a specified distance from the ground truth surface, providing a normalized measure of boundary alignment, with a higher value indicating better alignment. The NSD between two surfaces P and Q within a maximum tolerated distance τ is given by:

$$\text{NSD}(P, Q)^{(\tau)} = \frac{|S_P \cap \mathcal{B}_Q^{(\tau)}| + |S_Q \cap \mathcal{B}_P^{(\tau)}|}{|S_P| + |S_Q|} \tag{4}$$

where: $\mathcal{B}_Q^{(\tau)}$ is the border region of Q within the maximum tolerated surface distance, which defines the set of points that lie within a distance τ from the surface S_Q. A point $p \in S_P$ is considered to lie within this region if: $p \in \mathcal{B}_Q^{(\tau)} \iff d(p, S_Q) < \tau$. Similarly, $\mathcal{B}_P^{(\tau)}$ is the border region of P within the maximum tolerated distance τ. $|S_P \cap \mathcal{B}_Q^{(\tau)}|$ is the number of boundary points of P that lie within the border region of Q. $|S_Q \cap \mathcal{B}_P^{(\tau)}|$ is the number of boundary points of Q that lie within the border region of P. $|S_P|$ and $|S_Q|$ denote the total number of boundary points in P and Q, respectively. The NSD measures the proportion of boundary points of each

TABLE III
OVERVIEW OF ALL SUBMITTED METHODS ON BONBID-HIE 2023 LESION SEGMENTATION CHALLENGE. DETAILS OF EACH ALGORITHM'S DESCRIPTION CAN BE FOUND IN SUPPLEMENTARY FILE.

Team color	Participants	Brief description of algorithms
	civalab [46]	Integrated Swin-UNETR with random forest
	xleratorxlerator9 [47]	Decoder denoising self-pretraining and finetuning
	frimpz [48]	Swin-UNETR
	schlauglab [49]	Heavy augmentation with 3-D ResUNet
	rajroy	Swin-UNETR
	IWM [50]	3D ResUNet with heavy augmentation
	imad.toubal	Swin-UNETR
	UNetImage	Voxel specific logistic regression
	ngzvh	Swin-UNETR
	ashwin_dhakal	3D-UNet
	civa	3D-UNet
	arda.aydn [51]	SegResNet with Reciprocal Transformation
	SVCC [52]	nnUNet
	tiansong_philips	Swin-UNETR

surface that are within the maximum tolerated distance of the other surface's border region.

For the NSD metric, we used $\tau = 2$ mm, which aligns with the typical in-plane voxel spacing of our MRI data [20]. Specifically, the GE 1.5T scanner provided images with a resolution of $1.5 \times 1.5 \times (2.0\text{–}4.0)$ mm^3, while the SIEMENS 3T scanner offered a uniform resolution of $2 \times 2 \times 2$ mm^3. Given this range, we chose the 2.0 mm threshold to reflect the most common voxel dimension across datasets, consistent with standard practice and our imaging protocol in [20]. Moreover, the segmented lesions in our neonatal brain MRI are relatively small. It is a clinically acceptable boundary tolerance for this specific task. Any deviation beyond this distance would likely be considered a significant segmentation error. A τ value of 2 mm for boundary tolerance in neonatal brain MRI lesion segmentation is clinically justified because this matches the minimum voxel size used in clinical diffusion MRI for HIE [54]. Other non-BONBID-related independent neonatal lesion segmentation studies also reported an average surface distance of at 2 mm or slightly above as the central indicator of credible segmentation quality by atlas-based approach [55], conventional machine learning [56], [57] or deep learning [58], directly linking deviations beyond this range with significant segmentation error and reduced clinical utility.

Handling of empty segmentation masks. When the predicted mask is empty while the ground truth mask is non-empty, the MASD becomes undefined. Prior to evaluation, we reviewed all algorithms for instances of empty predicted masks: among the 14 algorithms, one produced a single empty mask, another produced three, and the remaining 12 each produced two. This indicates that all submitted algorithms occasionally generated empty predictions in this small and diffuse lesion segmentation task, even top performing methods. For these specific cases, a DICE score of zero and an NSD of zero were assigned. For MASD, we applied two computation strategies: (1) *Mean value substitution*: the missing value was replaced with the mean MASD computed from the respective algorithm's valid predictions across the test set (Reported in Table IV); and (2) *Max value substitution*: to penalize the empty prediction masks, the missing value was replaced with the maximum MASD observed from that algorithm's predictions across the test set (Reported in Table V). We chose this conservative strategy instead of imposing stronger penalties because the lesions are often small, and some patients may not present with visible lesions [20]; in such cases, empty outputs may represent clinically plausible scenarios.

Participation. Different algorithms (typically in articles by different first authors [30]) were allowed, even if the authors were from the same research group. Multiple submissions of the same algorithms, e.g., differences only in the parameter settings, were not allowed.

Rankings. The BONBID-HIE challenge employed a case-wise ranking method, which accounts for the significant variability in the complexity of patient cases. This ranking scheme has been successfully used in other challenges involving small and diffuse lesions, such as the subacute ischemic stroke lesion segmentation [21]. The evaluation process involved (1) computing the DICE, MASD, and NSD values for each case, (2) establishing each team's rank based on these metrics across all cases, and then (3) calculating the mean rank over all three evaluation measures to determine the final team rankings.

III. RESULTS

A. Submitted algorithms and leaderboard

During the challenge testing phase, 14 algorithms, software dockers, and method descriptions were submitted. All submitted dockers were evaluated on the hidden test set (N=44) on our challenge platform. Table III contains an overview of the methods submitted by the participating groups in the challenge (details are in the challenge proceedings [30] and challenge method supplementary file on our website github[4]). The submitted algorithms span a broad array of approaches leveraging

[4]https://github.com/baorina/BONBID-HIE-MICCAI2023

TABLE IV

(A) EVALUATION OF SUBMITTED METHODS ON HIE LESION SEGMENTATION USING DICE, MASD, AND NSD METRICS. THE TABLE RANKS 14 PARTICIPATING ALGORITHMS AND THE BASELINE. (B) LABEL FUSION USING MV, STAPLE, AND SBA ALGORITHMS ARE PRESENTED FOR THE TOP 2, 5, 10, AND ALL METHODS. GREYED-OUT ROWS INDICATE THE TOP-PERFORMING METHODS FOR EACH CORRESPONDING METRIC OR SUBGROUP.

(a) Ranking of 14 submitted algorithms.

Ranking	Participants	DICE (↑)	MASD (↓)	NSD (↑)
1	civalab	$62.2 \pm 24.4\%$	2.2 ± 2.5 mm	$75.6 \pm 24.1\%$
2	xleratorxlerator9	$62.3 \pm 23.9\%$	2.4 ± 2.6 mm	$74.9 \pm 23.6\%$
3	frimpz	$57.4 \pm 23.9\%$	2.7 ± 3.3 mm	$73.4 \pm 24.9\%$
4	schlauglab	$58.0 \pm 25.6\%$	2.6 ± 3.0 mm	$72.7 \pm 24.8\%$
5	rajroy	$57.1 \pm 24.1\%$	2.7 ± 3.4 mm	$73.2 \pm 25.1\%$
6	IWM	$57.7 \pm 25.1\%$	2.9 ± 3.6 mm	$72.2 \pm 25.2\%$
7	imad.toubal	$53.4 \pm 24.3\%$	2.5 ± 2.6 mm	$71.6 \pm 24.3\%$
8	UNeImage	$56.3 \pm 25.4\%$	3.0 ± 3.3 mm	$71.5 \pm 24.8\%$
9	ngzvh	$53.8 \pm 24.6\%$	3.0 ± 3.3 mm	$69.9 \pm 25.2\%$
10	ashwin_dhakal	$49.1 \pm 25.1\%$	3.4 ± 3.7 mm	$67.7 \pm 24.5\%$
11	civa	$50.0 \pm 26.3\%$	3.5 ± 3.4 mm	$67.8 \pm 23.7\%$
12	arda.aydn	$48.3 \pm 22.8\%$	3.5 ± 3.1 mm	$61.8 \pm 24.0\%$
13	punithakumar	$50.0 \pm 28.1\%$	5.2 ± 9.4 mm	$62.7 \pm 28.9\%$
14	tiansong_philips	$40.7 \pm 25.7\%$	4.7 ± 7.2 mm	$57.8 \pm 26.6\%$
Baseline	Z_{ADC} (-2)	$58.3 \pm 26.4\%$	3.2 ± 3.3 mm	$67.1 \pm 25.0\%$

(b) Accuracies by the consensus of algorithms.

Top 2	MV fusion [59]	$61.5 \pm 23.5\%$	2.4 ± 2.5 mm	$74.5 \pm 23.3\%$
	STAPLE fusion [60]	$61.8 \pm 24.2\%$	2.3 ± 2.4 mm	$75.3 \pm 23.9\%$
	SBA fusion [61]	$63.0 \pm 23.7\%$	2.3 ± 2.5 mm	$76.2 \pm 24.2\%$
Top 5	MV fusion	$62.9 \pm 23.8\%$	2.3 ± 2.9 mm	$76.5 \pm 24.7\%$
	STAPLE fusion	$60.7 \pm 23.4\%$	2.4 ± 2.7 mm	$75.0 \pm 23.8\%$
	SBA fusion	$61.7 \pm 24.7\%$	2.6 ± 3.3 mm	$74.1 \pm 26.3\%$
Top 10	MV fusion	$61.8 \pm 23.8\%$	2.4 ± 3.2 mm	$76.1 \pm 24.6\%$
	STAPLE fusion	$56.8 \pm 24.2\%$	2.6 ± 2.8 mm	$71.8 \pm 24.2\%$
	SBA fusion	$59.7 \pm 26.1\%$	3.0 ± 3.8 mm	$72.3 \pm 27.4\%$
all	MV fusion	$61.9 \pm 24.3\%$	2.5 ± 3.3 mm	$75.9 \pm 25.1\%$
	STAPLE fusion	$54.5 \pm 24.8\%$	2.7 ± 3.0 mm	$70.2 \pm 24.5\%$
	SBA fusion	$54.3 \pm 30.8\%$	4.1 ± 6.8 mm	$64.2 \pm 32.6\%$

anatomical information about HIE, data augmentation, training strategies, model architecture, and integration with traditional machine learning methods.

In summary, the top two methods exemplify different strategies for addressing the challenges of HIE lesion segmentation. Civalab's approach integrates Swin-UNETR [62] with traditional machine learning method Random Forest, focusing on refining segmentation accuracy. xleratorxlerator9's method emphasizes denoising and pretraining to enhance the quality of initial feature representations prior to fine-tuning for specific tasks. As indicated in the Table III, the majority of teams preferred advanced deep learning models, particularly Swin-UNETR and 3D-UNet variants, often paired with extensive data augmentation. Only the UNetImage team employed a traditional logistic regression model for lesion segmentation. The diversity of approaches reflects the ongoing experimentation and adaptation within the field, with some teams opting for traditional machine learning methods while others explored the potential of state-of-the-art deep learning methods.

Table IV summarizes the leaderboard rankings of all submitted algorithms. The ranking of the participating teams demonstrates a gradual improvement in the performance of the ranked approaches. For further evaluation of the impact of empty predictions, Table V reports the MASD results obtained when the maximum MASD from each algorithm's valid predictions across the test set is used to represent the MASD for empty predictions. It is noteworthy that the variability in evaluation metrics across the teams does not differ significantly between any two sequentially ranked teams.

B. Winning Method

The Top 1 method *civalab* combined a Swin-UNETR with a random forest classifier [63]. In the first stage, the Swin-UNETR processed the 3D HIE images using two channels (ADC and Z_{ADC}) to generate a lesion probability map. The

TABLE V
PERFORMANCE OF ALL TEAMS AND Z_{ADC} ON MASD, WHERE EMPTY PREDICTIONS WERE REPLACED BY EACH ALGORITHM'S MAXIMUM PREDICTED MASD VALUE (MEAN ± SD).

ID	Team	MASD
1	civalab	2.7 ± 3.1 mm
2	xleratorxlerator9	2.9 ± 3.4 mm
3	frimpz	3.4 ± 4.8 mm
4	schlauglab	3.3 ± 4.1 mm
5	rajroy	3.5 ± 4.9 mm
6	IWM	3.5 ± 4.3 mm
7	imad.toubal	2.9 ± 3.4 mm
8	UNeImage	3.5 ± 4.2 mm
9	ngzvh	3.6 ± 4.2 mm
10	ashwin_dhakal	4.0 ± 4.7 mm
11	civa	3.8 ± 4.0 mm
12	arda.aydn	4.1 ± 4.1 mm
13	punithakumar	7.2 ± 13.0 mm
14	tiansong_philips	7.6 ± 12.9 mm
15	Z_{ADC}(-2)	3.9 ± 4.6 mm

second stage involved a local refinement strategy: this probability map, alongside the input channels, was segmented into 5×5 2D windows and fed into a random forest classifier for a more localized lesion probability prediction. This two-stage approach, along with the use of the random forest for local refinement, played a role in mitigating the overfitting tendencies often associated with large parameter networks like Swin-UNETR, leading to more accurate and robust segmentations. Additionally, their proposed log Hausdorff distance loss aimed to regularize the 3D anatomical shape of HIE regions and implicitly optimize surface distance metrics.

The second-ranked method *xleratorxlerator9* also employed a two-stage process, beginning with Label Aware Denoising Pretraining (LADP). In the first stage, a deep learning model was pretrained using a denoising method LADP, which strategically applied increasing levels of noise to regions surrounding lesion contours. This pretraining aimed to enable the model to learn more robust and relevant features specifically for distinguishing lesion from surrounding tissue. The second stage leveraged the learned representations from the pretraining phase for the downstream segmentation task. For complete method description, please refer to [47].

In summary, both top methods used a two-stage strategy, benefiting from the separation of feature learning and refinement. Civalab used a global-to-local prediction approach with a random forest, while xleratorxlerator9 applied lesion-focused denoising pretraining.

C. Impact of different lesion volumes

Table VI evaluates the performance of submitted methods on HIE lesion cases, categorized by lesion size: $< 1\%$, $[1\%, 5\%]$, and $> 5\%$ of the brain volume. Lesions smaller than 1% are particularly important as they account for over 50% of the cases in the HIE dataset, underscoring the need for algorithms capable of accurately detecting and segmenting these tiny, often diffuse, lesions. As shown in the table, current methods are more effective at segmenting larger lesions, which

TABLE VI
EVALUATION OF SUBMITTED METHODS ON VARIOUS HIE LESION CASES.

Participants	DICE (↑)	MASD(↓)	NSD(↑)
Lesion $< 1\%$ of the whole brain volume			
civalab	$50.0 \pm 24.9\%$	3.4 ± 3.0 mm	$66.9 \pm 28.6\%$
xleratorxlerator9	$49.0 \pm 23.6\%$	3.9 ± 3.0 mm	$64.4 \pm 27.3\%$
frimpz	$44.5 \pm 22.8\%$	4.2 ± 4.2 mm	$62.5 \pm 29.2\%$
schlauglab	$43.7 \pm 24.1\%$	4.0 ± 3.6 mm	$62.1 \pm 28.6\%$
rajroy	$42.8 \pm 22.1\%$	4.3 ± 4.3 mm	$61.3 \pm 29.3\%$
IWM	$41.8 \pm 22.4\%$	4.9 ± 4.3 mm	$59.1 \pm 28.8\%$
imad.toubal	$38.3 \pm 21.0\%$	3.8 ± 3.1 mm	$59.3 \pm 27.6\%$
UNeImage	$40.8 \pm 22.4\%$	4.9 ± 3.8 mm	$59.0 \pm 27.4\%$
ngzvh	$38.4 \pm 20.9\%$	4.8 ± 3.9 mm	$56.9 \pm 28.1\%$
ashwin_dhakal	$31.7 \pm 17.9\%$	5.5 ± 4.3 mm	$53.6 \pm 25.5\%$
civa	$35.2 \pm 22.6\%$	5.5 ± 3.9 mm	$55.7 \pm 24.8\%$
arda.aydn	$37.8 \pm 21.2\%$	4.8 ± 3.9 mm	$56.4 \pm 27.2\%$
punithakumar	$35.8 \pm 26.9\%$	8.7 ± 12.3 mm	$51.0 \pm 32.7\%$
tiansong_philips	$24.0 \pm 17.6\%$	7.3 ± 9.3 mm	$45.4 \pm 28.4\%$
Z_{ADC} (-2)	$42.7 \pm 24.3\%$	5.2 ± 3.7 mm	$52.8 \pm 25.6\%$
Lesion $[1\%, 5\%]$ of the whole brain volume			
civalab	$66.8 \pm 15.8\%$	1.6 ± 0.9 mm	$80.1 \pm 15.2\%$
xleratorxlerator9	$69.3 \pm 14.2\%$	1.4 ± 0.7 mm	$82.5 \pm 11.8\%$
frimpz	$60.3 \pm 13.9\%$	1.7 ± 0.9 mm	$79.0 \pm 13.0\%$
schlauglab	$63.5 \pm 16.7\%$	2.0 ± 1.3 mm	$77.6 \pm 15.8\%$
rajroy	$62.2 \pm 13.4\%$	1.6 ± 0.8 mm	$80.5 \pm 11.0\%$
IWM	$66.0 \pm 12.2\%$	1.5 ± 0.6 mm	$81.2 \pm 9.5\%$
imad.toubal	$58.7 \pm 12.7\%$	1.6 ± 0.7 mm	$79.1 \pm 11.4\%$
UNeImage	$64.5 \pm 16.3\%$	1.7 ± 0.7 mm	$80.0 \pm 13.3\%$
ngzvh	$59.3 \pm 14.4\%$	1.8 ± 0.7 mm	$77.6 \pm 12.9\%$
ashwin_dhakal	$55.1 \pm 12.6\%$	1.9 ± 0.8 mm	$76.1 \pm 12.5\%$
civa	$52.1 \pm 17.3\%$	2.2 ± 1.0 mm	$72.6 \pm 15.3\%$
arda.aydn	$55.4 \pm 17.1\%$	2.5 ± 1.1 mm	$69.8 \pm 16.3\%$
punithakumar	$56.3 \pm 18.7\%$	2.7 ± 2.3 mm	$70.6 \pm 17.7\%$
tiansong_philips	$47.4 \pm 16.8\%$	3.2 ± 2.3 mm	$64.7 \pm 18.1\%$
Z_{ADC} (-2)	$67.4 \pm 17.9\%$	2.1 ± 0.8 mm	$76.9 \pm 12.7\%$
Lesion $> 5\%$ of the whole brain volume			
civalab	$83.4 \pm 12.0\%$	0.7 ± 0.4 mm	$89.6 \pm 9.0\%$
xleratorxlerator9	$83.3 \pm 12.7\%$	0.7 ± 0.5 mm	$88.8 \pm 12.0\%$
frimpz	$82.3 \pm 11.8\%$	0.7 ± 0.4 mm	$90.5 \pm 7.9\%$
schlauglab	$83.0 \pm 12.2\%$	0.7 ± 0.3 mm	$90.0 \pm 6.7\%$
rajroy	$82.4 \pm 11.8\%$	0.7 ± 0.4 mm	$90.5 \pm 7.9\%$
IWM	$82.7 \pm 14.9\%$	0.6 ± 0.4 mm	$90.3 \pm 8.9\%$
imad.toubal	$80.5 \pm 12.9\%$	0.7 ± 0.3 mm	$89.8 \pm 6.7\%$
UNeImage	$80.7 \pm 14.4\%$	0.7 ± 0.5 mm	$88.7 \pm 11.0\%$
ngzvh	$81.0 \pm 12.8\%$	0.8 ± 0.3 mm	$89.2 \pm 7.2\%$
ashwin_dhakal	$80.3 \pm 13.8\%$	0.8 ± 0.4 mm	$88.8 \pm 8.1\%$
civa	$79.8 \pm 13.5\%$	0.8 ± 0.4 mm	$88.5 \pm 7.6\%$
arda.aydn	$62.8 \pm 21.1\%$	2.1 ± 1.1 mm	$64.0 \pm 21.2\%$
punithakumar	$73.7 \pm 20.2\%$	1.3 ± 0.8 mm	$78.8 \pm 17.6\%$
tiansong_philips	$69.1 \pm 20.5\%$	1.4 ± 0.6 mm	$76.8 \pm 14.1\%$
Z_{ADC} (-2)	$81.6 \pm 14.5\%$	0.8 ± 0.5 mm	$87.0 \pm 12.6\%$

are generally easier to delineate, as evidenced by higher DICE scores, reaching up to 0.83 for lesions greater than 5%. In contrast, performance drops notably for smaller lesions, where even the top methods achieve DICE scores only in the 0.60–0.69 range. This trend is further illstrated that the best-performing methods for lesions smaller than 1% managed to achieve a DICE score of just around 0.50. The MASD and NSD metrics also reflect the increased complexity and variability associated with segmenting these small, diffuse lesions, indicating lower precision and accuracy. These results show the significant challenge of segmenting HIE lesions and highlight the pressing need for more advanced and refined algorithms capable of handling these difficult cases.

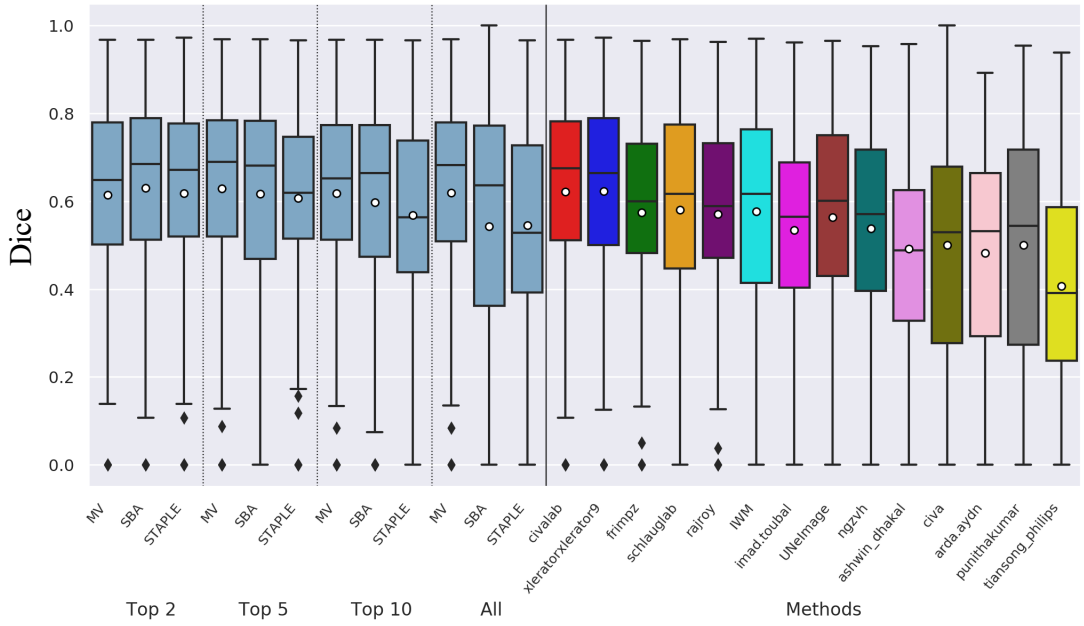

Fig. 5. Boxplot comparison of Dice scores across different label fusion strategies (MV, SBA, STAPLE) and individual team methods. The boxplots show the distribution of Dice scores for the top 2, top 5, top 10, and all algorithms, followed by each individual method.

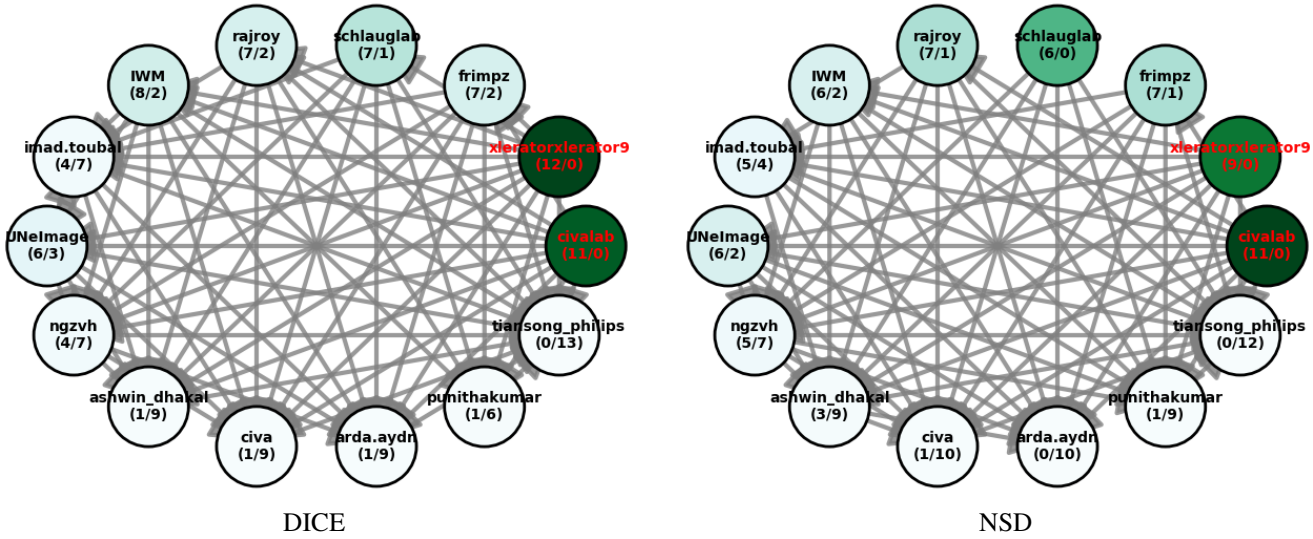

Fig. 6. Comparative analysis of 14 participating methods in terms of DICE and NSD metrics, evaluated using a one-sided Wilcoxon signed-rank test. Each node represents a method, with edges indicating statistically significant superiority of the originating method over the destination method. The number of outgoing and incoming edges, denoted as ($\#out/\#in$), reflects the relative strength of each method, with higher out-degrees and lower in-degrees signaling stronger performance. Node color saturation corresponds to this performance ratio, with more saturated colors highlighting more robust methods. Methods with identical edge counts exhibit comparable performance.

D. Statistical analysis

In order to assess potential statistically significant performance differences across teams, we also performed a pairwise comparison. Each pair of methods was compared using the one-sided Wilcoxon signed-rank test [64], a robust nonparametric test designed to determine whether one method consistently outperforms another. As shown in Figure 6, in the DICE metric evaluation, "civalab" and "xleratorxlerator9" emerged as the top-performing methods. Statistical analysis revealed no significant difference between these two methods ($p > 0.05$), indicating comparable performance. Furthermore, both methods demonstrated superiority over all 11 remaining methods. "Schlauglab" and "frimpz" were identified as the next highest-ranking methods, with "schlauglab" statistically outperforming seven other methods and "frimpz" showing superiority over the same number of competitors. However, both were statistically inferior to "civalab" and "xleratorxlerator9." In the NSD metric analysis, "civalab" and "xleratorxlerator9" continued to dominate, while "schlauglab" was found to statistically outperform six other teams. Similarly, "frimpz" outperformed seven methods, with only one method showing better performance. This analysis highlights the dominance of

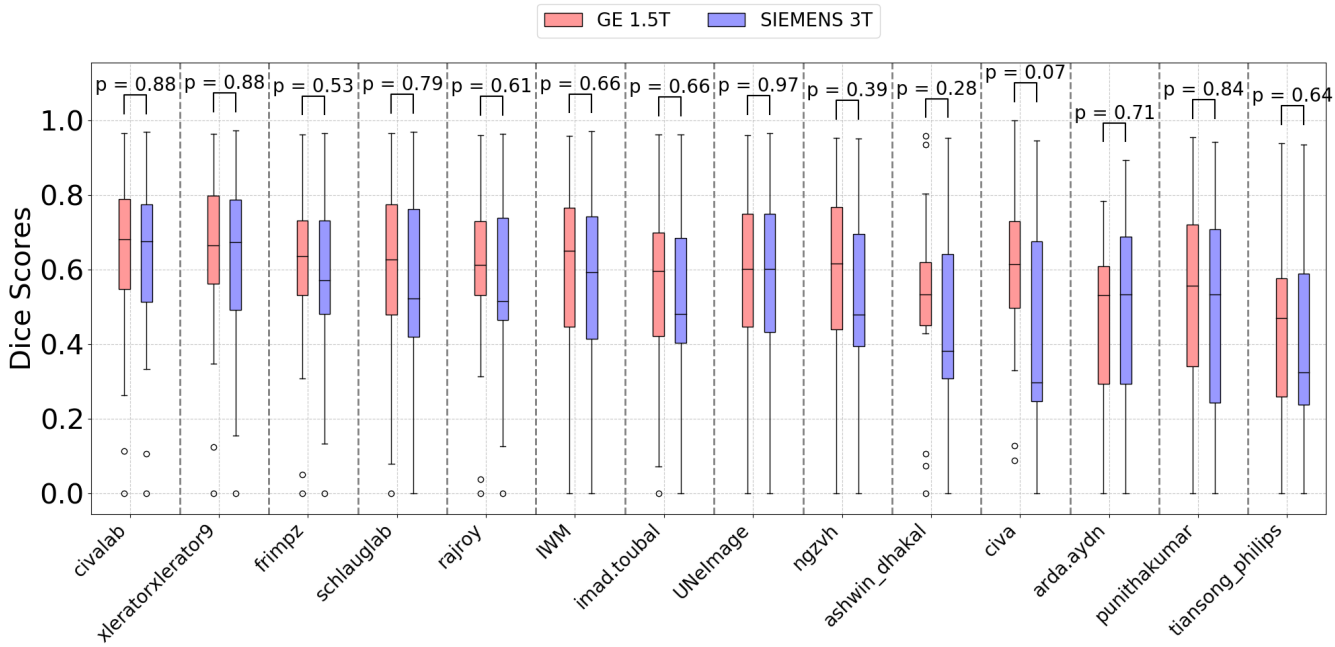

Fig. 7. Boxplot of DICE scores across different teams' algorithms on two types of scanners: GE 1.5T (red) and SIEMENS 3T (blue). The x-axis represents different teams, while the y-axis shows the DICE score values.

"civalab" and "xleratorxlerator9" in DICE and the strength of "schlauglab" and "frimpz" in NSD, confirming their high rankings and consistent performance in the initial leaderboard.

E. Impact of different field strength

Table II presents the distribution of patients in different lesion sizes across 1.5T GE and 3T SIEMENS scanners. Although efforts were made to split the test set with similar distributions, the SIEMENS scanner data contains nearly twice as many small and diffuse cases compared to the GE scanner data due to inherent data distribution. Cases acquired from different scanners can exhibit significant variations in appearance. A robust automatic HIE lesion segmentation method should effectively handle variations across scanners. To assess this, we evaluated each method's performance separately for each scanner type and conducted a Mann–Whitney U test [65] to compare the results between scanners for each algorithm.

Figure 7 illustrates the performance variability of different teams' algorithms, as measured by DICE scores. These methods show slightly better performance on GE data than on SIEMENS, as indicated by higher median DICE scores for several approaches, though this difference is not consistent across all methods. The boxplots demonstrate that the top two teams, "civalab" and "xleratorxlerator9", achieve stable performance across both scanners, with similar median scores and interquartile ranges. Other algorithms display greater variability, possibly due to challenges in detecting small, diffuse lesions that are more common in 3T SIEMENS scans. However, Mann–Whitney U test results indicate that these differences are not statistically significant ($p > 0.05$). Overall, the top two methods exhibit strong robustness across both 1.5T GE and 3T SIEMENS data, while the outliers in several methods suggest difficulties with specific lesion types or imaging conditions.

F. Combining the participants' results by label fusion

To assess whether label fusion can enhance lesion segmentation results, we applied three typical algorithms: Majority Voting (MV) [59], which assigns a lesion label when the majority of algorithms concur; STAPLE [60], which calculates global weights for each algorithm to reduce the influence of less accurate algorithms; and Shape-Based Averaging (SBA) [61], which iteratively improves performance by excluding the least accurate algorithms compared to a weighted majority voting.

We conducted label fusion using the top 2, top 5, top 10, and all methods. Table IV (b) and Figure 5 illustrated that the SBA of the top 2 algorithms yielded the better DICE, MASD and NSD, offered marginal improvement over the top 2 methods in DICE. Additionally, when using MV, STAPLE, or SBA, the negative influence of multiple less accurate algorithms correlated segmentations resulted in lower accuracy compared to at least the two top-ranked algorithms. This observation, that ensemble results may not outperform the best single algorithm, aligns with findings in the segmentation of subacute stroke lesions [21], which are also small and diffuse.

G. Failure Cases

We categorize the failure cases of the top two algorithms and the majority voting results into five main categories, as illustrated in Figure 8: (1) Small and Diffuse Lesions: HIE cases often involve lesions affecting less than 1% of brain tissue, posing significant challenges due to extreme data imbalance; (2) Atypical Lesion Locations: lesions in less common regions are underrepresented in the training data,

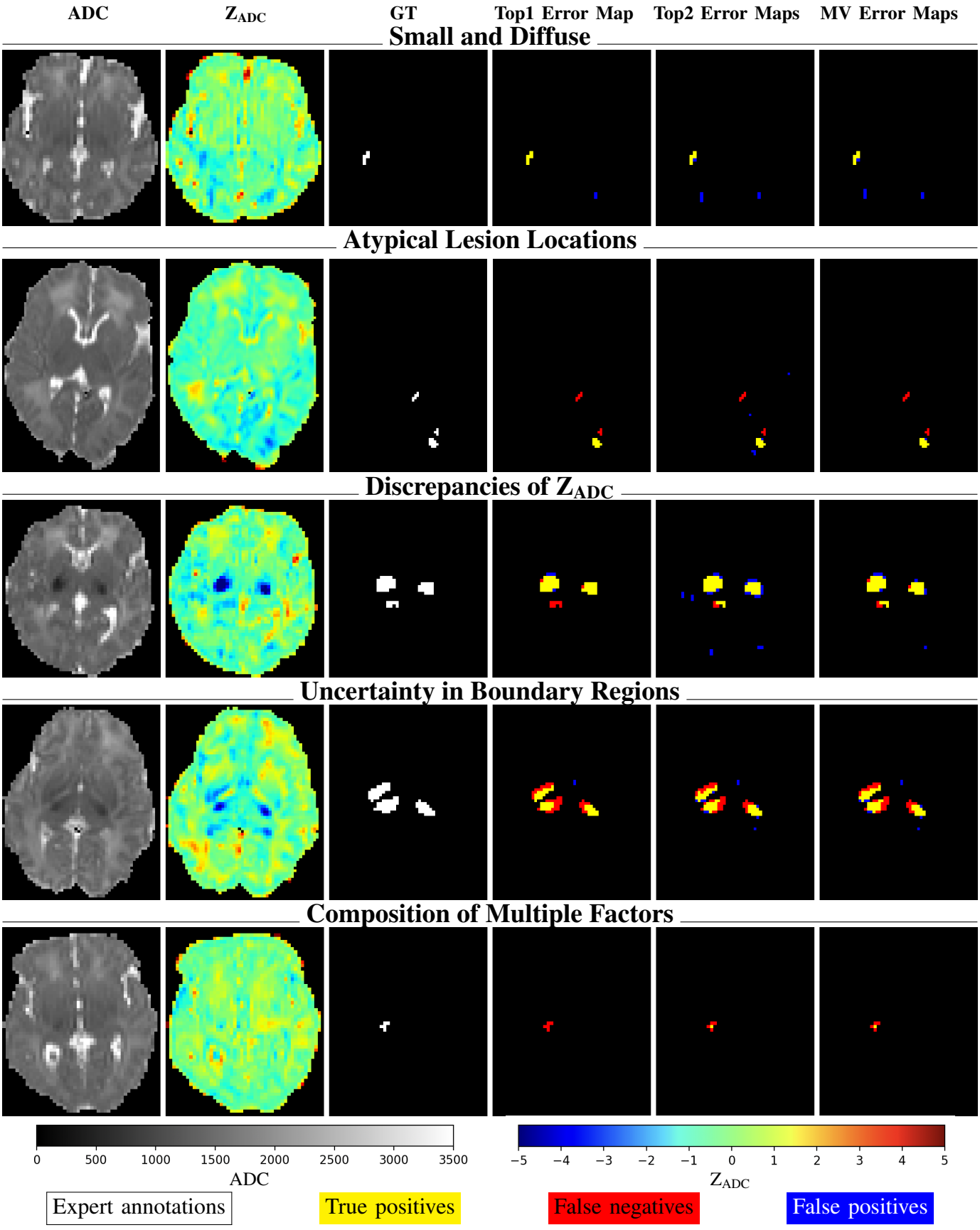

Fig. 8. Representative failure cases in HIE lesion segmentation. Each row illustrates a different challenge contributing to segmentation errors. Columns show the ADC map, Z_{ADC} map, ground truth (GT), and error maps from the top-1, top-2, and majority-vote (MV) methods. Error maps highlight true positives (yellow), false negatives (red), and false positives (blue), with expert annotations in green. These examples illustrate the complexity and ambiguity of neonatal HIE lesion segmentation and provide insights into common failure cases.

leading to suboptimal performance in these atypical cases. (3) Discrepancies of Z_{ADC}: variations in lesions can result in conflicting information between Z_{ADC} and ADC maps, leading to inaccurate segmentation inputs. (4) Uncertainty in Boundary Regions: ambiguous areas with inconsistent annotations, such as boundaries, making training more difficult and sometimes degrading segmentation performance. As shown in the figure, all methods struggle at boundary regions, where distinctions are challenging. (5) Composition of Multiple Factors: most algorithm failures stem from multiple concurrent issues.

IV. Discussion

Performance of segmentation algorithms on HIE. Despite advances in automated segmentating HIE lesions in our challenge, their performance and robustness still fall short of expectations (Overall DICE 0.62 and lesion $< 1\%$ DICE 0.5). Continuous improvement is expected through the expansion of training datasets to include larger and more diverse patient populations, along with innovations in training strategies and machine learning architectures. Future research should prioritize improving the robustness of automatic segmentation systems and addressing failure cases as discussed above, to achieve more reliable performance across varying lesion sizes in clinical HIE MRI settings.

Comparisons with Z_{ADC} maps. We provided a baseline method based on clinical knowledge rather than deep learning algorithms. This approach applies a straightforward thresholding of Z_{ADC} maps at a value of -2. Despite the lack of complex modeling or machine learning algorithms, this method ranked 3rd when compared with other machine learning-based methods submitted to the challenge (Table IV). This result highlights three important insights: (1) *Clinical Knowledge*: The baseline method, which is purely based on clinical comprehension of ADC maps, demonstrates that conventional, knowledge-driven approaches can still be highly competitive in medical imaging tasks. Although machine learning approaches frequently produce innovation and enhanced accuracy, the efficacy of this more straightforward strategy emphasizes the importance of established clinical concepts in achieving successful outcomes. (2) Interpretability versus Complexity: A primary advantage of this method is its interpretability. In contrast to deep learning models, which frequently function as "black boxes", such threshold-based method provides transparent and explicable criteria for its predictions. This transparency is a crucial benefit in medical contexts, where explainability is essential for clinical decision-making. (3) *Potential for Hybrid Approaches*: The success of this baseline method suggests that forthcoming research may gain from the integration of clinical expertise with machine learning methodologies. Hybrid models that integrate data-driven methodologies with domain-specific expertise may surpass solely machine learning-based methods while preserving a level of clinical interpretability.

Consensus of algorithms. Consensus among algorithms generally yields higher performance, as demonstrated in segmentation tasks for large focal lesion regions such as the Brain Tumor Segmentation (BraTS) challenge [22]. In these cases, consensus methods often outperform the single best algorithm, whether in the segmentation of the whole tumor region, tumor core, or active tumor area. However, this improvement does not always hold for small and diffuse lesions, such as in the ISLES-SISS (Subacute Ischemic Stroke Segmentation) task [21]. In such cases, consensus among algorithms may not outperform the top-performing single method, which we have also observed in the HIE lesion segmentation task.

Evaluation metrics. While we reported separate results for DICE, MASD, and NSD, and based our overall final ranking on these scores, these metrics may not adequately capture the performance of algorithms on small and diffuse lesions [53], [66]. However, there is no predefined metric that effectively evaluates small and diffuse lesions. DICE coefficient, although widely used, can disproportionately penalize slight mismatches in small regions, leading to lower scores even when the segmentation is clinically acceptable. Similarly, MASD and NSD, which rely on surface distance calculations, may struggle to accurately evaluate tiny or diffuse lesions where the boundary is not well-defined. Recent work on mesh-based metrics [67] compute distances directly in the mesh domain, which uses marching cubes or flying edges to generate surface meshes and accounts for boundary element sizes when computing distances. This approach may help reduce discretization artifacts and evaluate distance calculations more faithfully to reflect the underlying geometry of the segmentation. Incorporating mesh-based metrics in future evaluations may provide more reliable assessments for lesions with poorly defined boundaries.

Besides, the NSD metric is influenced by the choice of τ value [66]. Currently, we are using $\tau = 2$ mm, although it matches the most common in-plane voxel size in our dataset, it also reflects a clinically acceptable tolerance for small neonatal brain lesions. However, when the tolerance parameter τ approaches the the same order of magnitude as the image resolution, the distances between predicted and reference boundaries become discretized [66]. In the current study, only a single set of expert annotations was available, so we could not determine τ based on inter-observer variability. Our ongoing work involves collecting multiple expert annotations from a larger multi-center dataset. In future work, we will determine τ by considering both the clinical tolerance for our lesion segmentation task and expert-to-expert surface distances, following the approach of Nikolov *et al.* [66]. We will also investigate how alternative τ values influence metric sensitivity and ranking stability.

Inter-observer variance. One limitation of this study is the absence of multiple independent annotations for each subject. BONBID-HIE is the first public dataset for HIE lesion segmentation contains only a single annotation per subject, which was derived from a multi-expert consensus [20]. The lack of independent, multi-expert annotations hinders the ability to quantify intra- and inter-reader variability. Our ongoing multi-center work, spanning 21 sites and over 500 cases, incorporate multi-rater segmentations or consensus-driven labeling pipelines to address this limitation.

Clinical knowledge guided algorithms. Most participating teams approached HIE lesion segmentation from an artificial intelligence perspective, utilizing state-of-the-art methods tai-

lored for this task. Looking ahead, we encourage the formation of interdisciplinary teams that combine expertise from both the clinical and AI fields. By doing so, we aim to foster the development of algorithms that are guided by clinical knowledge, ensuring that the solutions are more aligned with clinical explanations, needs and practices. Besides, it remains to be elucidated how sex, age at MRI scan (in days), and other clinical factors will influence or contribute to lesion segmentation accuracy for HIE cases.

Beyond segmentation and toward outcome prediction. In the second challenge, we further enhanced the clinical relevance of HIE lesion segmentation by integrating it with HIE outcome prediction. The BONBID-HIE 2023 challenge has concluded, but the benchmarking platform remains available for future research. A follow-up challenge, held in association with MICCAI 2024, introduced an outcome prediction track and a modified segmentation task, along with additional unannotated data to support unsupervised and semi-supervised approaches. The results of this challenge will be presented in a separate article. We have introduced a new track focused on using MRIs to predict 2-year neurocognitive outcomes. Moreover, MRI data alone may not be sufficient for accurate outcome prediction. Therefore, in our upcoming challenges over the next few years, we plan to release clinical information and encourage the research community to incorporate this alongside MRI data for outcome prediction tasks.

V. Summary and Conclusion

Fourteen state-of-the-art approaches provided valuable insights and highlighted the persistent challenges in accurately segmenting HIE lesions. Our findings demonstrate that while segmentation of HIE lesions in MRI is feasible, the results are still suboptimal, particularly for lesions occupying less than 1% of brain volume, where DICE scores averaged only 0.50. Overall, the average DICE score across all cases was 0.62, underscoring the inherent difficulty of this task. Looking ahead to the next iteration of BONBID-HIE, we plan to expand the dataset by including data from additional sites and to introduce a neurocognitive outcome prediction task, linking MRI findings with practical clinical outcomes. Ultimately, we hope to facilitate the integration of AI methods with MRI and clinical data into clinical practice and to promote more interdisciplinary research in this field.

Acknowledgements

This work was funded, in part, by the Harvard Medical School and Boston Children's Hospital through Early Career Development Fellowship, Thrasher Research Fund Early Career Awards, NIH R21NS121735, R61NS126792, and R03HD104891.

We thank all the participants of the challenge and workshop for their valuable contributions and engagement.

References

[1] Rebecca J Weiss, Sara V Bates, Ya'nan Song, Yue Zhang, Emily M Herzberg, Yih-Chieh Chen, Maryann Gong, Isabel Chien, Lily Zhang, Shawn N Murphy, et al. Mining multi-site clinical data to develop machine learning MRI biomarkers: application to neonatal hypoxic ischemic encephalopathy. *Journal of translational medicine*, 17(1):1–16, 2019.

[2] Ernest M Graham, Kristy A Ruis, Adam L Hartman, Frances J Northington, and Harold E Fox. A systematic review of the role of intrapartum hypoxia-ischemia in the causation of neonatal encephalopathy. *American Journal of Obstetrics and Gynecology*, 199(6):587–595, 2008.

[3] Anne CC Lee, Naoko Kozuki, Hannah Blencowe, Theo Vos, Adil Bahalim, Gary L Darmstadt, Susan Niermeyer, Matthew Ellis, Nicola J Robertson, Simon Cousens, et al. Intrapartum-related neonatal encephalopathy incidence and impairment at regional and global levels for 2010 with trends from 1990. *Pediatric Research*, 74(1):50–72, 2013.

[4] Seetha Shankaran, Abbot R Laptook, Richard A Ehrenkranz, Jon E Tyson, Scott A McDonald, Edward F Donovan, Avroy A Fanaroff, W Kenneth Poole, Linda L Wright, Rosemary D Higgins, et al. Whole-body hypothermia for neonates with hypoxic–ischemic encephalopathy. *New England Journal of Medicine*, 353(15):1574–1584, 2005.

[5] A David Edwards, Peter Brocklehurst, Alistair J Gunn, Henry Halliday, Edmund Juszczak, Malcolm Levene, Brenda Strohm, Marianne Thoresen, Andrew Whitelaw, and Denis Azzopardi. Neurological outcomes at 18 months of age after moderate hypothermia for perinatal hypoxic ischaemic encephalopathy: synthesis and meta-analysis of trial data. *Bmj*, 340, 2010.

[6] Denis V Azzopardi, Brenda Strohm, A David Edwards, Leigh Dyet, Henry L Halliday, Edmund Juszczak, Olga Kapellou, Malcolm Levene, Neil Marlow, Emma Porter, et al. Moderate hypothermia to treat perinatal asphyxial encephalopathy. *New England Journal of Medicine*, 361(14):1349–1358, 2009.

[7] Abbot R Laptook, Seetha Shankaran, Jon E Tyson, Breda Munoz, Edward F Bell, Ronald N Goldberg, Nehal A Parikh, Namasivayam Ambalavanan, Claudia Pedroza, Athina Pappas, et al. Effect of therapeutic hypothermia initiated after 6 hours of age on death or disability among newborns with hypoxic-ischemic encephalopathy: a randomized clinical trial. *Jama*, 318(16):1550–1560, 2017.

[8] Seetha Shankaran, Abbot R Laptook, Athina Pappas, Scott A McDonald, Abhik Das, Jon E Tyson, Brenda B Poindexter, Kurt Schibler, Edward F Bell, Roy J Heyne, et al. Effect of depth and duration of cooling on death or disability at age 18 months among neonates with hypoxic-ischemic encephalopathy: a randomized clinical trial. *Jama*, 318(1):57–67, 2017.

[9] Zulian Liu, Tengbin Xiong, and Catherine Meads. Clinical effectiveness of treatment with hyperbaric oxygen for neonatal hypoxic-ischaemic encephalopathy: systematic review of chinese literature. *Bmj*, 333(7564):374, 2006.

[10] Molly Potter, Ted Rosenkrantz, and R Holly Fitch. Behavioral and neuroanatomical outcomes in a rat model of preterm hypoxic-ischemic brain injury: effects of caffeine and hypothermia. *International Journal of Developmental Neuroscience*, 70:46–55, 2018.

[11] Antonio Nuñez-Ramiro, Isabel Benavente-Fernández, Eva Valverde, Malaika Cordeiro, Dorotea Blanco, Hector Boix, Fernando Cabañas, Mercedes Chaffanel, Belén Fernández-Colomer, Jose Ramón Fernández-Lorenzo, et al. Topiramate plus cooling for hypoxic-ischemic encephalopathy: a randomized, controlled, multicenter, double-blinded trial. *Neonatology*, 116(1):76–84, 2019.

[12] Shi-Peng Liang, Qian Chen, Yi-Bing Cheng, Ying-Ying Xue, and Hai-Jun Wang. Comparative effects of monosialoganglioside versus citicoline on apoptotic factor, neurological function and oxidative stress in newborns with hypoxic-ischemic encephalopathy. *Coll Physicians Surg Pak*, 29(4):324–327, 2019.

[13] C Michael Cotten, Amy P Murtha, Ronald N Goldberg, Chad A Grotegut, P Brian Smith, Ricki F Goldstein, Kimberley A Fisher, Kathryn E Gustafson, Barbara Waters-Pick, Geeta K Swamy, et al. Feasibility of autologous cord blood cells for infants with hypoxic-ischemic encephalopathy. *The Journal of pediatrics*, 164(5):973–979, 2014.

[14] Abbot R Laptook, Seetha Shankaran, Patrick Barnes, Nancy Rollins, Barbara T Do, Nehal A Parikh, Shannon Hamrick, Susan R Hintz, Jon E Tyson, Edward F Bell, et al. Limitations of conventional magnetic resonance imaging as a predictor of death or disability following neonatal hypoxic–ischemic encephalopathy in the late hypothermia trial. *The Journal of pediatrics*, 230:106–111, 2021.

[15] Seetha Shankaran, Scott A McDonald, Abbot R Laptook, Susan R Hintz, Patrick D Barnes, Abhik Das, Athina Pappas, Rosemary D Higgins, Richard A Ehrenkranz, Ronald N Goldberg, et al. Neonatal magnetic resonance imaging pattern of brain injury as a biomarker of childhood outcomes following a trial of hypothermia for neonatal hypoxic-ischemic encephalopathy. *The Journal of pediatrics*, 167(5):987–993, 2015.

[16] Seetha Shankaran, Patrick D Barnes, Susan R Hintz, Abbott R Laptook, Kristin M Zaterka-Baxter, Scott A McDonald, Richard A Ehrenkranz, Michele C Walsh, Jon E Tyson, Edward F Donovan, et al. Brain injury following trial of hypothermia for neonatal hypoxic–ischaemic

encephalopathy. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 97(6):F398–F404, 2012.

[17] Yi Li, Jessica L Wisnowski, Lina Chalak, Amit M Mathur, Robert C McKinstry, Genesis Licona, Dennis E Mayock, Taeun Chang, Krisa P Van Meurs, Tai-Wei Wu, et al. Mild hypoxic-ischemic encephalopathy (hie): timing and pattern of mri brain injury. *Pediatric research*, 92(6):1731–1736, 2022.

[18] Bo Lyun Lee, Dawn Gano, Elizabeth E Rogers, Duan Xu, Stephany Cox, A James Barkovich, Yi Li, Donna M Ferriero, and Hannah C Glass. Long-term cognitive outcomes in term newborns with watershed injury caused by neonatal encephalopathy. *Pediatric research*, 92(2):505–512, 2022.

[19] Lauren C Weeke, Floris Groenendaal, Kalyani Mudigonda, Mats Blennow, Maarten H Lequin, Linda C Meiners, Ingrid C van Haastert, Manon J Benders, Boubou Hallberg, and Linda S de Vries. A novel magnetic resonance imaging score predicts neurodevelopmental outcome after perinatal asphyxia and therapeutic hypothermia. *The Journal of pediatrics*, 192:33–40, 2018.

[20] Rina Bao, Ya'nan Song, Sara V Bates, Rebecca J Weiss, Anna N Foster, Camilo Jaimes, Susan Sotardi, Yue Zhang, Randy L Hirschtick, P Ellen Grant, et al. Boston neonatal brain injury data for hypoxic ischemic encephalopathy (bonbid-hie): I. mri and lesion labeling. *Scientific Data*, 12(1):53, 2025.

[21] Oskar Maier, Bjoern H Menze, Janina Von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017.

[22] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[23] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[24] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[25] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[26] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[27] Stefan Winzeck, Arsany Hakim, Richard McKinley, José AADSR Pinto, Victor Alves, Carlos Silva, Maxim Pisov, Egor Krivov, Mikhail Belyaev, Miguel Monteiro, et al. Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral mri. *Frontiers in neurology*, 9:679, 2018.

[28] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022.

[29] Rina Bao, Rebecca J Weiss, Sara V Bates, Ya'nan Song, Sheng He, Jingpeng Li, Alte Bjornerud, Randy L Hirschtick, P Ellen Grant, and Yangming Ou. Paradise: Personalized and regional adaptation for hie disease identification and segmentation. *Medical Image Analysis*, 102:103419, 2025.

[30] Rina Bao, Ellen Grant, Andrew Kirkpatrick, Juan Wachs, and Yangming Ou, editors. *AI for Brain Lesion Detection and Trauma Video Action Recognition*, volume 14567 of *Lecture Notes in Computer Science*, Cham, 2024. Springer. doi:10.1007/978-3-031-71626-3.

[31] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

[32] Yangming Ou, Lilla Zöllei, Xiao Da, Kallirroi Retzepi, Shawn N Murphy, Elizabeth R Gerstner, Bruce R Rosen, P Ellen Grant, Jayashree Kalpathy-Cramer, and Randy L Gollub. Field of view normalization in multi-site brain mri. *Neuroinformatics*, 16:431–444, 2018.

[33] Yangming Ou, Randy L Gollub, Kallirroi Retzepi, Nathaniel Reynolds, Rudolph Pienaar, Steve Pieper, Shawn N Murphy, P Ellen Grant, and Lilla Zöllei. Brain extraction in pediatric adc maps, toward characterizing neuro-development in multi-platform and multi-institution clinical images. *NeuroImage*, 122:246–261, 2015.

[34] Siddhartha Gaddamanugu, Omid Shafaat, Houman Sotoudeh, Amir Hossein Sarrami, Ali Rezaei, Zahra Saadatpour, and Aparna Singhal. Clinical applications of diffusion-weighted sequence in brain imaging: beyond stroke. *Neuroradiology*, 64(1):15–30, 2022.

[35] Russell K Lawrence and Terrie E Inder. Anatomic changes and imaging in assessing brain injury in the term infant. *Clinics in perinatology*, 35(4):679–693, 2008.

[36] RJ Vermeulen, WPF Fetter, L Hendrikx, PEM Van Schie, MS Van Der Knaap, and F Barkhof. Diffusion-weighted mri in severe neonatal hypoxic ischaemia: the white cerebrum. *Neuropediatrics*, 34(02):72–76, 2003.

[37] Lishya Liauw, Gerda van Wezel-Meijler, Sylvia Veen, MA Van Buchem, and Jeroen van der Grond. Do apparent diffusion coefficient measurements predict outcome in children with neonatal hypoxic-ischemic encephalopathy? *American Journal of Neuroradiology*, 30(2):264–270, 2009.

[38] Kirsten PN Forbes, James G Pipe, and Roger Bird. Neonatal hypoxic-ischemic encephalopathy: detection with diffusion-weighted mr imaging. *American journal of neuroradiology*, 21(8):1490–1496, 2000.

[39] Ronald L Wolf, Robert A Zimmerman, Robert Clancy, and John H Haselgrove. Quantitative apparent diffusion coefficient measurements in term neonates for early detection of hypoxic-ischemic brain injury: initial experience. *Radiology*, 218(3):825–833, 2001.

[40] Alaa A Sayed, Nagham NM Omar, Nafisa H Refaat, and Mohammed K Mahmoud. Role of diffusion-weighted magnetic resonance imaging in detection of neonatal hypoxic-ischemic encephalopathy. *Journal of Current Medical Research and Practice*, 5(1):115–120, 2020.

[41] Benjamin Y Huang and Mauricio Castillo. Hypoxic-ischemic brain injury: imaging findings from birth to adulthood. *Radiographics*, 28(2):417–439, 2008.

[42] M Cimpersek, NP Meglic, DP Panjan, A Skofljanec, and KS Popovic. The role of diffusion weighted imaging and magnetic resonance imaging scoring system in assessing the effectiveness of treatment with hypothermia in neonates with hypoxic-ischemic encephalopathy. *Neonat Pediatr Med*, 3(135):2, 2017.

[43] Yangming Ou, Lilla Zöllei, Kallirroi Retzepi, Victor Castro, Sara V Bates, Steve Pieper, Katherine P Andriole, Shawn N Murphy, Randy L Gollub, and Patricia Ellen Grant. Using clinically acquired mri to construct age-specific adc atlases: Quantifying spatiotemporal adc changes from birth to 6-year old. *Human Brain Mapping*, 38(6):3052–3068, 2017.

[44] Martha Douglas-Escobar and Michael D Weiss. Hypoxic-ischemic encephalopathy: a review for the clinician. *JAMA pediatrics*, 169(4):397–403, 2015.

[45] Ruili Wei, Chaonan Wang, Fangping He, Lirong Hong, Jie Zhang, Wangxiao Bao, Fangxia Meng, and Benyan Luo. Prediction of poor outcome after hypoxic-ischemic brain injury by diffusion-weighted imaging: A systematic review and meta-analysis. *Plos One*, 14(12):e0226295, 2019.

[46] Imad Eddine Toubal, Elham Soltani Kazemi, Gani Rahmon, Taci Kucukpinar, Mohamed Almansour, Mai-Lan Ho, and Kannappan Palaniappan. Fusion of deep and local features using random forests for neonatal hie segmentation. *AI FOR BRAIN LESION DETECTION AND TRAUMA VIDEO ACTION RECOGNITION: First Bonbid*, 14567:3, 2024.

[47] Dean Ninalga. Label aware denoising pretraining. *CMBES Proceedings*, 46, 2024.

[48] Elham Soltani Kazemi, Imad Eddine Toubal, Gani Rahmon, Taci Kucukpinar, Mohamed Almansour, Mai-Lan Ho, and Kannappan Palaniappan. Enhancing lesion segmentation in the bonbid-hie challenge: An ensemble strategy. *AI FOR BRAIN LESION DETECTION AND TRAUMA VIDEO ACTION RECOGNITION*, 14567:14, 2024.

[49] Chiranjeewee Prasad Koirala, Sovesh Mohapatra, and Gottfried Schlaug. An ensemble approach for segmentation of neonatal hie lesions. In *AI FOR BRAIN LESION DETECTION AND TRAUMA VIDEO ACTION RECOGNITION*, pages 23–27. Springer, 2023.

[50] Marek Wodzinski and Henning Müller. Improving segmentation of hypoxic ischemic encephalopathy lesions by heavy data augmentation: Contribution to the bonbid challenge. In *AI FOR BRAIN LESION*

DETECTION AND TRAUMA VIDEO ACTION RECOGNITION, pages 28–33. Springer, 2023.

[51] M Arda Aydın, Elvin Abdinli, and Gozde Unal. Segresnet based reciprocal transformation for bonbid-hie lesion segmentation. In *booktitle=AI FOR BRAIN LESION DETECTION AND TRAUMA VIDEO ACTION RECOGNITION,*, pages 39–44. Springer, 2023.

[52] Nazanin Tahmasebi and Kumaradevan Punithakumar. A deep neural network approach for the lesion segmentation from neonatal brain magnetic resonance imaging. In *AI FOR BRAIN LESION DETECTION AND TRAUMA VIDEO ACTION RECOGNITION*, pages 34–38. Springer, 2023.

[53] Annika Reinke, Minu D Tizabi, Carole H Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

[54] Kengo Onda, Eva Catenaccio, Jill Chotiyanonta, Raul Chavez-Valdez, Avner Meoded, Bruno P Soares, Aylin Tekes, Harisa Spahic, Sarah C Miller, Sarah-Jane Parker, et al. Development of a composite diffusion tensor imaging score correlating with short-term neurological status in neonatal hypoxic–ischemic encephalopathy. *Frontiers in Neuroscience*, 16:931360, 2022.

[55] Richard J Beare, Jian Chen, Claire E Kelly, Dimitrios Alexopoulos, Christopher D Smyser, Cynthia E Rogers, Wai Y Loh, Lillian G Matthews, Jeanie LY Cheong, Alicia J Spittle, et al. Neonatal brain tissue classification with morphological adaptation and unified segmentation. *Frontiers in neuroinformatics*, 10:12, 2016.

[56] Neil I Weisenfeld and Simon K Warfield. Automatic segmentation of newborn brain MRI. *Neuroimage*, 47(2):564–572, 2009.

[57] Xintian Yu, Yanjie Zhang, Robert E Lasky, Sushmita Datta, Nehal A Parikh, and Ponnada A Narayana. Comprehensive brain MRI segmentation in high risk preterm newborns. *PloS one*, 5(11):e13874, 2010.

[58] Leonie Richter and Ahmed E Fetit. Accurate segmentation of neonatal brain MRI with deep learning. *Frontiers in Neuroinformatics*, 16:1006532, 2022.

[59] Lei Xu, Adam Krzyzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics*, 22(3):418–435, 1992.

[60] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

[61] Torsten Rohlfing and Calvin R Maurer Jr. Shape-based averaging for combination of multiple segmentations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 838–845. Springer, 2005.

[62] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[63] Eddine Toubal, Elham Soltani Kazemi, Gani Rahmon, and Taci Kucukpinar. Fusion of deep and local features using random forests for neonatal hie segmentation. *AI for Brain Lesion Detection and Trauma Video Action Recognition*, pages 1–12, 2023.

[64] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[65] Michael P Fay and Michael A Proschan. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, 4:1, 2010.

[66] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[67] Gašper Podobnik and Tomaž Vrtovec. Metrics revolutions: Groundbreaking insights into the implementation of metrics for biomedical image segmentation. *arXiv preprint arXiv:2410.02630*, 2024.