Brain Extraction Methods in Neonatal Brain MRI and Their Effects on Intracranial Volumes

Tânia F. Vaz, Nuno Canto Moreira, Lena Hellström-Westas, Nima Naseh, Nuno Matela, and Hugo A. Ferreira

Abstract: Magnetic resonance imaging (MRI) plays an important role in assessing early brain development and injury in neonates. When using an automated volumetric analysis, brain tissue segmentation is necessary, preceded by brain extraction (BE) to remove non-brain tissue. BE remains challenging in neonatal brain MRI, and despite the existence of several methods, manual segmentation is still considered the gold standard. Therefore, the purpose of this study was to assess different BE methods in the MRI of preterm neonates and their effects on the estimation of intracranial volumes (ICVs). This study included twenty-two premature neonates (mean gestational age ± standard deviation: 28.4 ± 2.1 weeks) with MRI brain scans acquired at term, without detectable lesions or congenital conditions. Manual segmentation was performed for T2-weighted scans to establish reference brain masks. Four automated BE methods were used: Brain Extraction Tool (BET2); Simple Watershed Scalping (SWS); HD Brain Extraction Tool (HD-BET); and SynthStrip. Regarding segmentation metrics, HD-BET outperformed the other methods with median improvements of +0.031 (BET2), +0.002 (SWS), and +0.011 (SynthStrip) points for the dice coefficient; and −0.786 (BET2), −0.055 (SWS), and −0.124 (SynthStrip) mm for the mean surface distance. Regarding ICVs, SWS and HD-BET provided acceptable levels of agreement with manual segmentation, with mean differences of −1.42% and 2.59%, respectively.

Keywords: brain extraction; intracranial volume; neonatal MRI; segmentation; skull stripping

1. Introduction

Magnetic resonance imaging (MRI) plays an important role in assessing early brain development and injury in neonates through clinical interpretation and volumetric analyses (e.g., intracranial volume (ICV) and regional brain volume measurements) [1,2]. In this context, image acquisition and assessment pose unique challenges [2–4]: (1) fast imaging sequences, with the incomplete filling of the k-space, are typically used in order to avoid sedation and movement artifacts, but at the expense of image quality; (2) the use of two-dimensional (2D) sequences with thick slices to further decrease the acquisition time leads to partial volume effects which may compromise the assessment of the small brain structures being imaged; (3) the contrast-to-noise ratio (CNR) between grey and white matter may be lower if imaging sequences are not optimized regarding repetition and echo times; (4) if dedicated coils are not available, the signal-to-noise ratio (SNR) might also be compromised; (5) the appearance of the brain and regional structures differs from the adult brain and undergoes rapid changes with gestational age (GA) and postnatal maturation (due to evolving myelination, decreases in brain water content, and an increase in tissue density); (6) abnormalities may be related to specific perinatal pathologies that are not
typically observed in adults and usually evolve rapidly; and (7) image processing methods also vary across premature newborns, infants, and adults.

As an essential pre-processing step for brain tissue segmentation and image registration, brain extraction (BE) or skull stripping is necessary to remove non-brain tissue [5–7]. This is a critical step because inaccurate intracranial segmentation, e.g., unremoved non-brain tissues or incorrectly removed brain tissues, could result in the under- or over-estimation of brain volumes [5].

There are several methods for BE (i.e., manual, semi-automated, and automated algorithms), but manual segmentation is still considered the gold standard or ground truth for BE; despite being time consuming and having inter/intra-rater variability, they are often used to validate semi-automated and automatic methods [6–11]. These methods can be classified as follows: conventional methods (deformable surface-based, mathematical morphology, intensity, template, and hybrid-based models) as well as machine learning and deep learning methods [7,9,11]. The majority of the available methods are focused on the adult brain (e.g., BEaST [12], BEMA [13], BET [8], BET2 [14], BSE [15], CONSNet [16], DMBE [17], FSW [18], HD-BET [19], MASS [20], McStrip [21], ROBEX [22], MONSTR [23], SBA [24], SPECTRE [25], SynthStrip [26], and the 3D U-Net approach [28]), but some have been optimized or developed specifically for neonates (e.g., ALFA [29], iBEAT2 [30], LABEL [31], STAPLE [32], GCSP [10], HSS [33], AFSS [34], and the fuzzy object model-based fuzzy connectedness approach [35]).

BE remains challenging in neonatal brain MRI because of the low spatial resolution, low signal-to-noise ratio, low contrast-to-noise ratio, wide variation in intensity within tissues, small brain size, evolving shape, and motion artifacts when compared with adult brain images [5,6,11]. Additionally, most BE methods are designed to work with T1-weighted (T1w) MRI, but in neonates, these images lack contrast between brain tissues, so T2-weighted (T2w) images are most often used [1,2,36].

Therefore, the purpose of this study was to apply different BE methods in the MRI of preterm neonates and assess their performances, focusing on comparisons with manual segmentation as well as on the effect they have in estimating ICVs.

2. Materials and Methods

2.1. Subjects

The dataset used in this study consisted of pseudonymized MRI brain scans from 22 premature neonates (Table 1) without detectable brain lesions or congenital conditions, who had no severe motion artifacts and were scanned at term-equivalent age (TEA), which is considered the optimal timeframe for evaluating structural abnormalities that may have implications for long-term outcomes [37]. The MRI scans were performed as part of the clinical routine for very preterm infants at Uppsala University Hospital, a tertiary referral centre in Sweden, and parental consent was obtained for retrospective data analysis, as approved by the Human Research Ethical Committee of the Medical Faculty at Uppsala University (Dnr 2014/236).

Table 1. Premature neonates’ characteristics.

<table>
<thead>
<tr>
<th>All Infants (GA &lt; 32 Weeks)</th>
<th>Extremely Preterm (EP) (GA &lt; 28 Weeks)</th>
<th>Very Preterm (VP) (GA 28–31 Weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study population</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>Gestational age (GA) in weeks (mean ± standard deviation (range))</td>
<td>28.4 ± 2.1 (25–31)</td>
<td>26.0 ± 0.3 (25–27)</td>
</tr>
<tr>
<td>Gender (female, male)</td>
<td>13, 9</td>
<td>5, 3</td>
</tr>
</tbody>
</table>

2.2. MRI Protocol

All MRI scans were acquired on a Siemens Avanto 1.5 Tesla scanner (Siemens Medical Systems, Erlangen, Germany) at Uppsala University Hospital using a neonatal-adapted...
imaging protocol. To perform this study, 2D fast spin echo (FSE) T2w images were acquired with the following parameters: axial slice number = 33; slice thickness = 3 mm; gap distance factor 20% interleaved (spacing between slices = 3.6 mm); field-of-view (FOV) read/phase = 200 mm/100%; repetition time (TR) = 6520 ms; echo time (TE) = 103 ms; and flip angle = 120°.

2.3. Manual Segmentation for Brain Extraction

To create a reference brain mask for each image volume, a strictly defined, fully manual procedure was followed using the 3D Slicer (version 5.2.2) software [38]. The mask was drawn slice by slice in the axial plane, considering the cerebral contour [39], excluding non-brain structures, such as the neck, ears, eyes, scalp, and skull, with reference to brain atlases [40,41], using the original images without preprocessing as input. Each brain mask was also divided into stack of slices, creating three masks for each BE method, from the bottom to the top of the brain, to assess whether the methods produced different results depending on the anatomical region. The bottom mask—slices 1 to 11—covered the base of the skull, and part of brain stem and cerebellum (inferior limit: first slice, level of the spinal cord; superior limit: level of the pons, when the ocular globes and the lenses are visible). The middle mask—slices 12 to 22—included subcortical structures such as the basal ganglia (superior limit: below the level of the centrum semiovale). The top mask—slices 23 to 33—covered the centrum semiovale and the vertex (superior limit: top of the skull).

Rater 1 (T.F.V.) performed all segmentations, which were validated by another two experienced raters (rater 2—N.C.M. and rater 3—H.A.F.). To evaluate intra-rater reliability, rater 1 re-segmented six randomly chosen subjects from previously segmented data after one month. Inter-rater accuracy was assessed including segmentations from raters 2 and 3, which manually outlined six randomly selected subjects from the dataset.

2.4. Selection of Automated Brain Extraction Methods

The criteria for selection of the skull stripping methods were based on their public availability, citations, generalizability for a variety of contrasts, and use of state-of-the-art methods without published studies in neonates.

Four automated BE methods were used (Table 2): two conventional methods (Brain Extraction Tool (BET2) [14] and Simple Watershed Scalping (SWS) [42]) and two deep learning methods (HD Brain Extraction Tool (HD-BET) [19] and SynthStrip [26]). Again, each brain mask obtained for each method was also divided into stack of slices, following the same procedure as described in Section 2.3.

Finally, whereas the different BE methods can be used within toolboxes or as standalone programs, in this study, the respective toolboxes were used (second column of Table 2), with the exception of SynthStrip [26], which was used as a standalone program running in Docker [43] (the software versions can be found in Section 2.6).

<table>
<thead>
<tr>
<th>Brain Extraction Method</th>
<th>Software</th>
<th>Method Category [9]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain Extraction Tool (BET version 2.0, BET2) [8,14]</td>
<td>FSL (FMRIB Software Library) [44]</td>
<td>Deformable surface-based model</td>
</tr>
<tr>
<td>HD Brain Extraction Tool (HD-BET, version 1.0) [19]</td>
<td>Extension of 3D Slicer [38] or as an open-source standalone</td>
<td>Deep learning (based on U-Net architecture and its 3D derivatives)</td>
</tr>
<tr>
<td>SynthStrip (version 1.3) [26]</td>
<td>FreeSurfer [47] or as an open-source standalone</td>
<td>Deep learning (based on 3D U-Net)</td>
</tr>
</tbody>
</table>

BET2 [14] is based on BET [8], one of the most widely used methods for skull stripping, using the original model or several modified versions [6,7,9]. BET2 uses techniques such as...
intensity clamping, surface point detection, and mesh fitting to find the brain boundary in brain MRI [14]. Even though they do not have parameters optimized for neonates, they have been used in this context [42,48–54].

SWS is a preliminary version of a BE tool available within the Morphologically Adaptive Neonatal Tissue Segmentation (MANTiS) toolbox [42], based on watershed transform. It can be used before the tissue classification pipeline and has been used in several studies [55–65], although it was not used in the validation study of the toolbox (they used BET instead) as it was not available then.

HD-BET [19] relies on artificial neural networks and has outperformed six commonly used BE tools (BET [8], 3dSkullStrip [27], BSE [15], ROBEX [22], BEaST [12], and MONSETR [23]). The algorithm was trained with multi-sequence MRI from adults (precontrast T1w, postcontrast T1w, T2w, and fluid-attenuated inversion recovery (FLAIR)) with different MRI scanners and acquisition parameters and showed robustness even in the presence of pathology- or treatment-induced tissue alterations [19].

SynthStrip [26] builds on a solid foundation laid by prior studies of deep learning algorithms for BE, using a 3D U-Net convolutional architecture. It has generally produced highly accurate brain masks compared with other BE tools (BET [8], BEaST [12], DMBE [17], FSW [18], ROBEX [22], and 3dSkullStrip [27]). It is a flexible BE tool that can be deployed universally for a variety of brain images because it is agnostic to acquisition parameters, as it never samples any real data during training because a strategy for synthesizing diverse training data is applied. Its efficacy was demonstrated in a dataset, ranging from infants to adults, that included structural MRI (T1w, T2w, FLAIR, proton density weighted, MR angiography, computed tomography, and $^{18}$F-Fluorodeoxyglucose positron emission tomography.

HD-BET and SynthStrip are state-of-the-art BE methods reporting excellent results, and to the best of our knowledge, there have not been focused research efforts on their BE performance in MRI preterm infants scanned at TEA.

From these methods, only BET2 and SynthStrip allowed for variations in options, and we systematically investigated their performance to identify the parameters that provided more accurate BE results compared with those of manual segmentation, since there are no references to use in neonates. Based on the results (Appendix A), we set these parameters to apply to the dataset.

As a note, we also tested the iBEAT2 BE method [30]. Nonetheless, at the time of this study, this method had some limitations: (1) it did not provide whole-brain parcellation, so when removing the brain skull, the cerebellum was also removed, and the final output could not be used to measure the ICV; and (2), when inserting the data details for the method, the minimum accepted age was 1 month (currently, the method accepts data from 0-month-old neonates that would be considered at the TEA). These limitations, therefore, precluded the use of this method and its comparison to others.

2.5. Measurement of Intracranial Volumes

ICV is defined as the volume inside the cranium [39,66], and in this study, the ICVs were estimated in mL with 3D Slicer (version 5.2.2) [38] from the intracranial masks obtained from each BE method (manual, BET2, SWS, HD-BET, and SynthStrip).

2.6. Hardware and Software

Our research was performed on a MacBook Pro Retina (Apple Inc., Cupertino, CA, USA) with a 2.7 GHz Intel Core i5 CPU and 8 GB of RAM. The manual segmentation was performed using an external 28” 4K UHD monitor (Lenovo L28u-35, Lenovo Group Ltd., Beijing, China) and a graphics tablet (GAOMON S620, Gaomon Technology Co., Guangzhou, China).

Several software were used to perform BE and volume measurements, such as the following: 3D Slicer (version 5.2.2) [38]; Docker (version 4.18.0) [43]; FSL (version 6.0.6.4) [44]; MATLAB® R2023a (version 9.14) [46]; and SPM12 (revision 7771) [45]. The segmentation
metrics were computed in Python (version 3.11.2) [67], and all statistical analyses were performed with the IBM® SPSS® v27 software [68].

2.7. Evaluation Metrics and Statistical Analysis

2.7.1. Evaluation Metrics

Several quantitative validation metrics have been used for BE assessments [9–11,33,69,70] because each metric yields different information:

- **Voxel-overlap-based metrics:** dice coefficient (DC); Jaccard coefficient (JC); precision (Pr); sensitivity (Se); specificity (Sp); false positive rate (FPR); and false negative rate (FNR).
- **Volume-based metrics:** volume similarity (VS); root mean square error (RMSE); and coefficient of variation of the root mean square error (CVRMSE).
- **Surface distance-based metrics:** Hausdorff distance (HD); Hausdorff distance 95% percentile (HD95); mean surface distance (MSD); median surface distance (MDSD); and standard deviation surface distance (STDSD).

The metrics’ formulas and their interpretations are described in Appendix B, and their values were computed in Python (version 3.11.2) [67].

2.7.2. Statistical Analysis

Intra- and inter-rater agreement of the reference standards—brain masks that were manually segmented—were assessed by calculating DC, and the ICVs were calculated with the intraclass correlation coefficient (ICC). ICC estimates and their 95% confidence intervals (CIs) were calculated based on a two-way mixed-effects model with a mean rating of (k = 2), absolute agreement, and ranges from 0 to 1, with higher values indicating greater levels of reliability [71].

To test the assumption of normality of distribution and homogeneity of variance [71], Shapiro–Wilk test and Levene’s test were used when applicable, respectively.

The independent samples t-test [71] was used to compare the ICVs of different genders (female (F); male (M)) and gestational age categories (extremely preterm (EP): <28 weeks; very preterm (VP): 28–31 weeks).

The segmentation assessment between BE methods was determined using voxel-overlap-based metrics, surface-distance-based metrics, and VS, and general differences were found using the Friedman test [71]. The differences between bottom, middle, and top masks were assessed using the DC and Friedman test. For post hoc comparisons, Bonferroni-adjusted significance tests were used for pairwise comparisons [71]. Agreement between ICVs extracted from MRI using manual and automated segmentation methods was reported using Bland–Altman analysis [71–73], root mean square error (RMSE), and coefficient of variation of the root mean square error (CVRMSE) [71].

All statistical analyses were performed with a significance value of 5%, using the IBM® SPSS® v27 software [68].

3. Results

3.1. Reliability Assessment of Manual Segmentation

Both the intra- and inter-rater agreement of the manual brain masks indicated a high level of overlap within and between raters. The intra-rater DC was on average 0.989 (range: 0.982–0.996) for rater 1, DC = 0.985 (range: 0.981–0.987) for rater 2, and DC = 0.984 (range: 0.983–0.985) for rater 3. The inter-rater DC results were on average 0.988 (range: 0.986–0.991) (rater 1 vs. 2) and 0.986 (range: 0.983–0.987) (rater 1 vs. 3). There were good to excellent levels of intra-rater reliability between the ICVs manually segmented by rater 1, and the average ICC measured (95% CI) was 0.996 (0.845–0.999). The inter-rater reliability was excellent, with the ICC of rater 1 vs. 2 = 0.999 with a 95% CI of 0.954–1.000, and the ICC of rater 1 vs. 3 = 0.996 with a 95% CI of 0.906–1.000.
3.2. Volume Differences between Gender and Gestational Age

The estimated ICVs manually showed similar results (Figure 1) between genders (ICV_M = 458.2 ± 55.8 mL; ICV_F = 466.3 ± 47.9 mL) and gestational age categories (ICV_EP = 459.4 ± 54.2 mL; ICV_VP = 465.1 ± 49.7 mL). According to the independent samples t-test, the differences between the ICVs showed no statistically significant differences between females and males (t(20) = −0.364; p = 0.719; Cohen’s d = −0.158) and between EP and VP neonates (t(20) = −0.251; p = 0.804; Cohen’s d = −0.111). Based on these results and the small effect sizes (d < 0.2) [71], all of the analyses were performed without separating the dataset into smaller subgroups.

![Figure 1. ICV measurements by gender (a) and gestational age (b).](image)

3.3. Comparison of Manual vs. Automated Brain Extraction Methods

3.3.1. Segmentation Metrics

According to Tables 3 and 4, generally, the best segmentation metrics were obtained when comparing the manual segmentation with HD-BET followed by SWS, with median DC values of 0.968 and 0.966, as well as MSD values of 0.596 mm and 0.651 mm, respectively (Figure A2a,k in Appendix C). Regarding VS, despite the smaller volume size difference given by SynthStrip (−0.083%), it has a wider IQR (4.718%) than those of the other methods (BET2, 2.798%; SWS, 2.902%; HD-BET, 2.687%). SynthStrip did not show statistically significant differences from SWS (|Z| = 0.455; p = 1.000) and HD-BET (|Z| = 1.000; p = 0.061), which had a tendency to underestimate volume (−2.033%) and overestimate volume (2.609%), respectively. All of our violin plots are available in Appendix C. Table 5 shows the DC obtained for each brain mask subset (bottom, middle, and top slices) with the different BE methods. HD-BET generally performed well in all slices (DCBottom = 0.944, DCMiddle = 0.979, and DCTop = 0.962), but in the middle, SWS was slightly better (DCMiddle = 0.983). The performance of all BE methods was superior in the middle slices of the brain MRI.

![Figure 1. ICV measurements by gender (a) and gestational age (b).](image)

Table 3. Voxel-overlap-based metrics statistics (median and interquartile range (IQR)) of four automated BE methods (BET2, SWS, HD-BET, and SynthStrip).

<table>
<thead>
<tr>
<th></th>
<th>DC</th>
<th>JC</th>
<th>Pr</th>
<th>Se</th>
<th>Sp</th>
<th>FPR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BET2</td>
<td>0.937 [0.933; 0.941]</td>
<td>0.881 [0.874; 0.889]</td>
<td>0.906 [0.896; 0.912]</td>
<td>0.971 [0.959; 0.980]</td>
<td>0.989 [0.988; 0.990]</td>
<td>0.011 [0.010; 0.012]</td>
<td>0.029 [0.020; 0.041]</td>
</tr>
<tr>
<td>SWS</td>
<td>0.966 [0.963; 0.969]</td>
<td>0.934 [0.929; 0.939]</td>
<td>0.977 [0.961; 0.981]</td>
<td>0.959 [0.954; 0.962]</td>
<td>0.998 [0.996; 0.998]</td>
<td>0.003 [0.002; 0.005]</td>
<td>0.041 [0.038; 0.046]</td>
</tr>
<tr>
<td>HD-BET</td>
<td>0.968 [0.965; 0.971]</td>
<td>0.938 [0.932; 0.944]</td>
<td>0.958 [0.948; 0.962]</td>
<td>0.983 [0.974; 0.987]</td>
<td>0.995 [0.994; 0.996]</td>
<td>0.005 [0.004; 0.006]</td>
<td>0.017 [0.013; 0.026]</td>
</tr>
<tr>
<td>SynthStrip</td>
<td>0.957 [0.952; 0.961]</td>
<td>0.917 [0.909; 0.925]</td>
<td>0.958 [0.944; 0.964]</td>
<td>0.961 [0.947; 0.969]</td>
<td>0.996 [0.994; 0.996]</td>
<td>0.005 [0.004; 0.006]</td>
<td>0.039 [0.031; 0.053]</td>
</tr>
</tbody>
</table>

Abbreviations: DC—Dice coefficient; JC—Jaccard coefficient; Pr—Precision; Se—Sensitivity; Sp—Specificity; FPR—False positive rate; FNR—False negative rate.
Table 4. Volume-based metrics and surface-distance-based metrics statistics (median and interquartile range (IQR)) of four automated BE methods (BET2, SWS, HD-BET, and SynthStrip).

<table>
<thead>
<tr>
<th></th>
<th>Median [IQR] of Volume-Based Metrics</th>
<th>Median [IQR] of Surface-Distance-Based Metrics (In mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VS (%)</td>
<td>HD95</td>
</tr>
<tr>
<td>BET2</td>
<td>7.185 [5.51, 9.31]</td>
<td>3.600 [3.600, 4.48]</td>
</tr>
<tr>
<td>SWS</td>
<td>[−2.03]</td>
<td>3.400 [3.12, 3.60]</td>
</tr>
<tr>
<td>HD-BET</td>
<td>2.609 [1.32, 4.00]</td>
<td>3.156 [2.50, 3.60]</td>
</tr>
<tr>
<td>SynthStrip</td>
<td>0.083 [−1.77, 2.94]</td>
<td>3.600 [3.12, 3.60]</td>
</tr>
</tbody>
</table>

Abbreviations: VS—Volume similarity; HD—Hausdorff distance; HD95—Hausdorff distance 95% percentile; MSD—Mean surface distance; MDSD—Median surface distance; STDSD—Standard deviation surface distance.

Table 5. Dice coefficient statistics (median and interquartile range (IQR)) of automated BE methods (BET2, SWS, HD-BET, and SynthStrip) for each brain mask subset (bottom, middle, and top slices).

<table>
<thead>
<tr>
<th></th>
<th>Median [IQR] of Dice Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bottom Mask (Slices 1 to 11)</td>
</tr>
<tr>
<td>BET2</td>
<td>0.874 [0.857, 0.883]</td>
</tr>
<tr>
<td>SWS</td>
<td>0.897 [0.877, 0.905]</td>
</tr>
<tr>
<td>HD-BET</td>
<td>0.944 [0.935, 0.947]</td>
</tr>
<tr>
<td>SynthStrip</td>
<td>0.926 [0.913, 0.930]</td>
</tr>
</tbody>
</table>

We identified general differences between the methods in all segmentation metrics (using the Friedman test with \( p < 0.001 \)), with Kendall’s W coefficient of concordance ranging from 0.548 to 0.808 across the different metrics, representing a large effect size \[71\], showing the strength or the meaningfulness between the manual and automated segmentation metrics. Concerning the DC results obtained for the bottom, middle, and top brain masks, statistically significant differences (using the Friedman test with \( p < 0.001 \)) were also identified. Regarding the post hoc pairwise comparisons (Table 6), statistically significant differences were identified for DC and JC between the methods (using the Bonferroni correction with \( p < 0.001 \)), except for the comparison of HD-BET and SWS (\(|Z| = 0.273; p = 1.000\)). However, if we analyse the surface-based metrics, in regards to HD95, MSD, and MDSD, there were no statistically significant differences between HD-BET and SWS, between SWS and SynthStrip, or between HD-BET and SynthStrip (except for MSD with \(|Z| = 1.045; p = 0.043\)). The VS showed no differences between HD-BET and SynthStrip (\(|Z| = 1.000; p = 0.061\)) or SWS and SynthStrip (\(|Z| = 0.455; p = 1.000\)). The post hoc pairwise comparisons of the brain mask subset (Table 7) showed variable results between the methods and slices, with an emphasis on non-statistically significant differences between HD-BET and SWS in the middle (\(|Z| = 1.868; p = 0.370\)) and top slices (\(|Z| = 0.117; p = 1.000\)), as well as between HD-BET and SynthStrip in the bottom (\(|Z| = 2.335; p = 0.117\)) and top (\(|Z| = 1.985; p = 0.283\)) slices. Three slices (one from each subset) overlapped with the brain masks are shown in Figure 2 as examples.

Table 6. Pairwise comparisons of segmentation metrics of the automated BE methods. For each comparison, we have reported the absolute value of the Z-statistics (\(|Z|\)) and the Bonferroni-adjusted \( p \)-value (adj. \( p \)).
Table 6. Cont.

<table>
<thead>
<tr>
<th>Surface-distance-based metrics</th>
<th>HD-BET vs. BET2</th>
<th>HD-BET vs. SWS</th>
<th>HD-BET vs. SynthStrip</th>
<th>SWS vs. BET2</th>
<th>SWS vs. SynthStrip</th>
<th>SynthStrip vs. BET2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume-based metrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VS</td>
<td>1.182</td>
<td>0.014</td>
<td>1.455</td>
<td>0.001</td>
<td>1.000</td>
<td>0.061</td>
</tr>
<tr>
<td>HD</td>
<td>2.568</td>
<td>0.000</td>
<td>1.727</td>
<td>0.000</td>
<td>0.614</td>
<td>0.690</td>
</tr>
<tr>
<td>HD95</td>
<td>2.318</td>
<td>0.000</td>
<td>0.545</td>
<td>0.967</td>
<td>0.682</td>
<td>0.479</td>
</tr>
<tr>
<td>MSD</td>
<td>2.545</td>
<td>0.000</td>
<td>0.591</td>
<td>0.774</td>
<td>1.045</td>
<td>0.043</td>
</tr>
<tr>
<td>MDSD</td>
<td>2.159</td>
<td>0.000</td>
<td>0.068</td>
<td>1.000</td>
<td>0.409</td>
<td>1.000</td>
</tr>
<tr>
<td>STDS</td>
<td>2.364</td>
<td>0.000</td>
<td>1.318</td>
<td>0.004</td>
<td>0.500</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Abbreviations: DC—Dice coefficient; JC—Jaccard coefficient; Pr—Precision; Se—Sensitivity; Sp—Specificity; FPR—False positive rate; FNR—False negative rate; VS—Volume similarity; HD—Hausdorff distance; HD95—Hausdorff distance 95% percentile; MSD—Mean surface distance; MDSD—Median surface distance; STDS—Standard deviation surface distance.

Table 7. Pairwise comparisons of DC for each brain mask subset and method. For each comparison, we have reported the absolute value of the Z-statistics (|Z|) and the Bonferroni-adjusted p-value (adj. p).

<table>
<thead>
<tr>
<th></th>
<th>HD-BET vs. BET2</th>
<th>HD-BET vs. SWS</th>
<th>HD-BET vs. SynthStrip</th>
<th>SWS vs. BET2</th>
<th>SWS vs. SynthStrip</th>
<th>SynthStrip vs. BET2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom mask (slices 1 to 11)</td>
<td>7.123</td>
<td>0.000</td>
<td>5.488</td>
<td>0.000</td>
<td>2.335</td>
<td>0.117</td>
</tr>
<tr>
<td>Middle mask (slices 12 to 22)</td>
<td>5.021</td>
<td>0.000</td>
<td>1.868</td>
<td>0.370</td>
<td>2.919</td>
<td>0.021</td>
</tr>
<tr>
<td>Top mask (slices 23 to 33)</td>
<td>5.839</td>
<td>0.000</td>
<td>0.117</td>
<td>1.000</td>
<td>1.985</td>
<td>0.283</td>
</tr>
</tbody>
</table>

Figure 2. TEA brain T2w MRI of a preterm neonate (born at 29 weeks GA) in three different axial slices (a) (from bottom to top: slice numbers 9, 17, and 26) and corresponding overlapping brain masks (in yellow) from each BE method: (b) Manual, (c) BET2, (d) SWS, (e) HD-BET, and (f) SynthStrip.

3.3.2. Intracranial Volume Estimation

Table 8 shows the ICV of the 22 premature babies scanned at TEA, estimated with the manual and automated methods. The Bland–Altman plots illustrate the agreement between the ICVs extracted from the MRI using automated segmentation methods against the reference (Figure 3). Ideally, the BE methods being compared would yield identical references (95% CI) or bias were as follows: 32.4 (12.8 to 52.0) mL for SWS vs. BET2, 39.3 (20.5 to 58.1) mL for BET2 vs. SWS, 10.3 (1.4 to 19.2) mL for HD-BET vs. SWS, and 9.5 (1.1 to 17.9) mL for SynthStrip vs. HD-BET.

Instead, we would expect to observe a random distribution of ICV differences clustered around zero, indicating an unbiased pattern of disagreement [71–73]. By examining the spread of differences both above and below zero, we can determine the extent to which the
disagreement between the methods is acceptable when replacing manual for automated BE methods. If the majority of values consistently fall above/below zero, it suggests that one method consistently produces higher/lower measurements than the other. This pattern indicates that the disagreement is not random but rather systematic [71–73]. Furthermore, by analysing the data, we can also identify whether distinct patterns of disagreement exist among subjects with lower or higher ICVs [71–73].

Table 8. Descriptive statistics of intracranial volumes (mL) obtained with different brain extraction methods.

<table>
<thead>
<tr>
<th>Brain Extraction Method</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>463.0</td>
<td>50.1</td>
<td>388.7</td>
<td>575.6</td>
<td>440.8</td>
<td>485.3</td>
</tr>
<tr>
<td>BET2</td>
<td>495.4</td>
<td>47.0</td>
<td>423.2</td>
<td>610.8</td>
<td>474.6</td>
<td>516.3</td>
</tr>
<tr>
<td>SWS</td>
<td>456.5</td>
<td>52.2</td>
<td>377.1</td>
<td>571.2</td>
<td>433.4</td>
<td>479.6</td>
</tr>
<tr>
<td>HD-BET</td>
<td>475.2</td>
<td>46.0</td>
<td>403.9</td>
<td>581.2</td>
<td>454.8</td>
<td>495.6</td>
</tr>
<tr>
<td>SynthStrip</td>
<td>464.8</td>
<td>40.2</td>
<td>395.7</td>
<td>554.4</td>
<td>447.0</td>
<td>482.6</td>
</tr>
</tbody>
</table>

![Graph](image)

**Figure 3.** Bland–Altman analysis of the mean of (x-axis) and the difference between (y-axis) the automated BE methods and manually segmented ICVs. (a) BET2—Manual, (b) SWS—Manual, (c) HD-BET—Manual, and (d) SynthStrip—Manual. The full lines (blue, red, green, and purple) indicate the mean difference, the dotted lines (blue, red, green, and purple) indicate upper and lower limits of agreement (±1.96 standard deviations), the thinner dotted lines in grey represent zero (no difference), and the linear regression line is shown in black.

The ICV mean differences (95% CI) or bias were as follows: 32.4 (12.8 to 52.0) mL for BET2 – Manual; −6.5 (−23.0 to 10.0) mL for SWS – Manual; 12.1 (−1.2 to 25.4) mL for HD-BET—Manual; and 1.8 (−25.4 to 28.9) mL for SynthStrip—Manual. These results
can be visually perceived in some slices of Figure 2, with masks exceeding the brain area (e.g., BET2 and HD-BET) or the opposite (e.g., SWS and SynthStrip), as identified in the Bland–Altman analysis.

Most differences in the ICVs from the Bland–Altman analysis (Figure 3) were within the limits of agreement, with a range both above and below the mean difference; however, the differences were not always evenly distributed as ideally expected, so a linear regression analysis \[71,73\] was performed to assess the presence of proportional bias. The regression analysis was the statistical procedure used for examining the predictive relationship among the difference in ICVs between each automated and manual BE method (dependent—criterion—variable) and the mean ICV estimated by those methods (independent—predictor—variable). The results for the ICV mean differences of HD-BET—Manual (R = 0.608; \(p = 0.003\)) and SynthStrip—Manual (R = 0.719; \(p < 0.001\)) showed heteroscedasticity. However, the homoscedasticity of BET2—Manual (R = 0.317; \(p = 0.151\)) and SWS—Manual (R = 0.240; \(p = 0.283\)) did not indicate the same tendency of error with the increase in ICV. The correlational trend line of HD-BET demonstrates that for certain higher ICVs, the magnitude of positive differences tends to slightly decrease, approaching that of manual segmentation (in which the differences are closer to zero). Conversely, the regression line of SynthStrip reflects a tendency for the negative differences to increase for higher ICVs, showing a larger variability in the ICV estimations. Converting the difference between methods to percentages results in a mean bias of 6.76% for BET2; −1.42% for SWS; 2.59% for HD-BET; and 0.38% for SynthStrip. The RMSE and CVRMSE values for each automated method relative to the reference were as follows: 33.8 mL (7.31%) for BET2; 10.5 mL (2.27%) for SWS; 13.8 mL (2.99%) for HD-BET; and 13.6 mL (2.95%) for SynthStrip.

### 3.4. Computation Time

Each scan took roughly 3 h to segment manually. The average processing time (in seconds) for each automated BE method was as follows: 2.20 ± 0.15 (BET2); 4.95 ± 1.33 (SWS); 1288.40 ± 12.37 (HD-BET); and 266.97 ± 89.24 (SynthStrip).

### 4. Discussion

This study assessed different BE methods for MRI scans of preterm neonates and their effects on ICVs, showing that automated methods have a similar level of accuracy to that of manual skull stripping in T2w MRI. The study’s prerequisites were met regarding the ground truth that was created for BE, because both the intra- and inter-rater showed a high level of agreement (with average DC and ICC values above 0.980), suggesting that the full manual segmentation performed was a reliable method for establishing reference brain masks (in agreement with the findings of other studies \[29,31\]). Additionally, the dataset was tested as a whole, since the ICV estimation did not show significant differences in terms of gender and gestational age.

Based on the segmentation metrics results, we can interpret the different BE methods (Manual, BET2, SWS, HD-BET, and SynthStrip). According to the voxel-overlap-based metrics, HD-BET followed by SWS obtained the highest results for median DC (0.968 (IQR: 0.965–0.971) and 0.966 (IQR: 0.963–0.969)) and JC (0.938 and 0.934) values, respectively. A similar deduction can be made from the DC assessment for each brain mask subset, considering that the segmentation performance varies between them, due to anatomical differences across slices. The lower segmentation accuracy in the bottom slices might be related to the presence of different brain tissues (the cerebrum, cerebellum, and brain stem), vascular structures, air-filled cavities, and the intricate internal structure of the skull bones. The top slices cover a larger area of the brain, where it merges with the surrounding blood vessels, exterior skull bones, and scalp, challenging the segmentation of the brain from non-brain adjacent structures. HD-BET outperforms the other tested BE methods with median improvements of +0.031 (BET2), +0.002 (SWS), and +0.011 (SynthStrip) points for DC; −0.786 (BET2), −0.055 (SWS), and −0.124 (SynthStrip) mm for MSD; and −0.44 mm for HD95. Similar results were reported with HD-BET by Isensee et al. \[19\] with different
adult MRI datasets and sequences, yielding a median DC of 0.976 (IQR, 0.970–0.980) on T1w, 0.969 (IQR, 0.961–0.974) on cT1w, 0.964 (IQR, 0.952–0.970) on FLAIR, and 0.961 (IQR, 0.952–0.967) on T2w sequences. Since SWS was not validated as a BE tool by Beare et al. [42], there are not studies comparing these results, despite their similarity to results from manual segmentation. However, the same authors proposed a BE method for T1w MRI also based on the watershed transform and validated it in human and macaque brains, showing a mean ± standard deviation DC in a pediatric cohort of 0.962 ± 0.005 [74], which is very similar to the results obtained in our study. When demonstrating the efficacy of SynthStrip, Hoopes et al. [26] included a small subset of 10 infants born full-term with T1w images acquired between 0 and 18 months, presenting similarities to our results, with a mean DC of 0.961 ± 0.014. Regarding surface-distance-based metrics, HD-BET performed well with a median HD95 of 3.2 mm (IQR: 2.5–3.6 mm), followed by SWS and SynthStrip, which also achieved good results (3.6 mm; IQR: 3.1–3.6 mm). Similar results were reported with HD-BET by Isensee et al. [19], yielding a median HD95 of 2.7 mm (IQR: 2.2–3.3 mm) on T1w, 3.2 mm (IQR: 2.8–4.1 mm) on cT1w, 4.2 mm (IQR: 3.4–5.0 mm) on FLAIR, and 4.4 mm (IQR: 3.9–5.0 mm) on T2w sequences. BET2’s overestimation of the brain boundary has also been previously reported by Smith [8]. Regarding SynthStrip, our HD95 and MSD results are better than those reported with an infant subset by Hoopes et al. [26] (38.7 ± 22.2 mm and 1.0 ± 0.3 mm, respectively) and might be related to the range of ages (0 to 18 months) included. In another study that evaluated twelve BE methods in a neonatal brain MRI dataset (preterms scanned at TEA), Serag et. al. [29] showed that the developed ALFA method [29] had a better performance than our tested methods, with a mean DC of 0.989 and an HD of 3.4 mm (for the T2w images). We intended to include this method in our study; however, it was no longer accessible, and therefore could not be tested with our dataset. Ten of the other tested methods [29] performed worse than HD-BET and SWS, namely, 3DSS (DC = 0.922), BET (DC = 0.792), BSE (DC = 0.714), LABEL (DC = 0.935), ROBEX (DC = 0.910), MV (DC = 0.951), STAPLE (DC = 0.948), SBA (DC = 0.960), BW (DC = 0.774), and BEaST (DC = 0.939). We also obtained different results when optimizing BET parameters, as Serag et. al. [29] showed a DC of 0.792, whereas in our study, it was 0.937, which is a massive improvement that suggests our parameters were better suited. The BE accuracy within a neonatal group (0.8 ± 0.3 months old) using the LABEL [31] method developed by Shi et. al. showed an average JC of 0.948 and their method performed better when compared with other BE methods (BET (JC = 0.920), BSE (JC = 0.884), ROBEX (JC = 0.886), GCUT (JC = 0.916), Majority voting (JC = 0.911), and STAPLE (JC = 0.934)). The LABEL results are slightly higher than those of HD-BET and SWS, but this could be due to the age of the dataset, since this method has a tendency for the JC to increase with the age of the sample [31]. Finally, the HSS method developed by Péporté et. al. [33] showed worse similarity metrics (DC = 0.954 and JC = 0.913) than HD-BET, SWS, and SynthStrip, as well as the other BE methods (FSL, BrainSuite, MRIcroN, and SPM8) they tested in their dataset.

Regarding ICV estimation, our Bland–Altman analysis indicated good agreement between some automated and manual BE methods, with the identification of patterns of disagreement that can be used as correction factors. BET2 consistently estimates higher ICVs than manual segmentation; this pattern is less prominent with HD-BET and with a tendency to approach to minor differences when the ICVs increase. On the other hand, SWS mostly estimates lower ICVs than manual segmentation. SynthStrip shows the largest variability, with a tendency to overestimate at lower ICVs and underestimate at higher ICVs. The ICV difference discrepancies between our study and the infant’s dataset by Hoopes et al. [26] (7.4 ± 3.5%) might be related to the wide variation in brain size between 0 and 18 months, the range of ages included in the test subset of SynthStrip. The error between automated and manual ICVs is small (<3%) for all the methods, except for BET2, indicating good agreement between the segmentation methods. However, the error obtained for BET2 was similar to that previously reported by Smith [8] when developing this widely used BE method; nonetheless, special awareness should be given when using this BE method in neonates due to the impact of a 7% measurement error when studying such small
brains. SWS and HD-BET provide a level of agreement that is likely to be acceptable for experimental and clinical applications due to the smaller ICV mean differences (−6.5 mL (−1.42%) and 12.1 mL (2.59%)), median VS (2.61% and −2.03%), RMSE (10.5 mL and 13.8 mL), and CVRMSE (2.27% and 2.99%) values. This means that if the ICVs estimated from the automated BE methods are corrected by a factor of +1.42% (SWS) or −2.59% (HD-BET), it should be close to the real ICV with a small estimation error (±1.13% for SWS and ±1.49% for HD-BET). The estimates might be considered suitable considering the neonatal volume variation per week. According to Dimitrova et al. [75], 6.1% per week is a typical ICV developmental change during the neonatal period in a term-born dataset (37–45 weeks). Moreover, between 32 and 42 weeks, Mewes et al. [76] reported a brain growth of 11.2% per week. Furthermore, Gui et al. [77] and Zacharia et al. [78] depicted an average ICV growth rate of 27.2 mL and 24.0 mL per week in cohorts of premature infants that were assessed at birth and later at TEA, respectively.

Skull stripping, as a preprocessing step, should be fast, providing results that are as accurate as those of manual segmentation, which nonetheless is extremely time consuming when compared with automated methods [6–11]. Herein, we have observed that conventional automated BE methods perform faster than deep-learning-based methods, in agreement with the computation times described by other authors [9,30,53]. However, this limitation of the latter methods could be overcome by running them in GPU mode instead of CPU mode (e.g., HD-BET takes approximately 30 s in GPU mode [19]).

This study achieves the following: (1) it determines the applicability of state-of-the-art BE methods, such as HD-BET and SynthStrip, in MRI preterm infants scanned at TEA; (2) validates SWS with manual segmentation; (3) optimizes the parameters for BET2 and SynthStrip to apply in neonates; and (4) culminates in a proposal of correction factors for BE automated methods.

The current study also had some limitations, namely the dataset used for this study was small. This is because for testing the outputs of BE methods, it is recommended (1) to include the full manual segmentation as a reference, which is a time-consuming process and which is the main reason why the majority of similar studies have small sample sizes (e.g., several studies [10,12,15,16,22,29,30,33,35] reported small datasets/subgroups ranging from 5 to 22 subjects); and (2) to test subjects without detectable brain lesions or congenital conditions (and no severe motion artifacts) first [29,31,34–36,42], criteria which greatly limit the available datasets, as in our case, most infants had brain abnormalities.

The potential implications of these restrictions are described below:

1. The influence of our sample size on our Bland–Altman analysis [79] is difficult to determine due to the unknown clinically acceptable agreement limits for ICV measurements of preterm neonates at term. This analysis indicates the likely range of true mean differences in the specific clinical population, considering the potential for random variation [71–73]. Nevertheless, the findings from a small sample size are less likely to capture the characteristics and variability of the overall population, mainly when statistical inferences are applied; therefore, caution is needed when generalizing conclusions.

2. The presence of brain lesions and congenital condition or motion artifacts might lead to difficulties in accurately delineating the border between the brain and the skull, depending on their severity, location (e.g., near the border of the skull or affecting skull integrity and brain tissue characteristics), and BE method category [9]. When using BET2 [8,14], which is a deformable surface-based model approach, the presence of brain pathology and motion can pose challenges in the evolution of the mesh used to approximate the brain surface, primarily due to intensity and shape variations. However, the BE methods based on deep learning techniques are known to mitigate this constraint [9]. For instance, Isensee et al. [19] stated that HD-BET showed a robust performance in the presence of adult brain pathology- or treatment-induced tissue alterations. Additionally, Hoopes et al. [26] developed SynthStrip to be agnostic to MRI contrasts and acquisition schemes, and it was validated for age, health, resolution,
and imaging modality factors. So, it is expected that deep learning methods behave similarly in a pathological neonatal context. Furthermore, motion artifacts can be present, especially in older infants, as it can be challenging to maintain their sleep during scanning [80]. Fortunately, since the typical protocols used in neonatal brain MRI (e.g., low magnetic field strengths and fast imaging sequences) do not provide high-resolution images, they are less sensitive to motion artifacts [81]. Nonetheless, special attention should be given when considering pathological cases with motion artifacts, because the performance of BE methods may vary considerably.

Concerning future work, it is recommended to expand the neonate’s dataset (with a larger sample size to increase the confidence of the findings) and perform BE in the presence of pathology, image artifacts, and different MRI protocols. Another interesting approach for further studies could be to assess regional brain volumes using pre-processed MRI scans of brains extracted with different methods.

In conclusion, HD-BET and SWS provide the best overall BE performances and can be considered, amongst the tested methods and considering our study’s limitations, the most suitable BE methods in the T2w MRI of preterm neonates scanned at TEA.

Author Contributions: Conceptualization, T.F.V. and H.A.F.; methodology, T.F.V. and H.A.F.; software, T.F.V.; validation, H.A.F. and N.C.M.; formal analysis, T.F.V. and H.A.F.; investigation, T.F.V.; resources, T.F.V., N.C.M., L.H.-W. and N.N.; data curation, T.F.V.; writing—original draft preparation, T.F.V.; writing—review and editing, T.F.V., H.A.F. and N.M.; visualization, T.F.V.; supervision, H.A.F., N.M. and N.C.M.; project administration, T.F.V., H.A.F., N.M. and N.C.M.; funding acquisition, T.F.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Fundação para a Ciência e a Tecnologia (FCT), under the IBEB Strategic Program UIDB/00645/2020 (https://doi.org/10.54499/UIDB/00645/2020), and Bolsa de Investigação para Doutoramento (2022.12483.BD).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Human Research Ethical Committee of the Medical Faculty at Uppsala University (protocol code 2014/236/1 and date of approval 24 February 2015, with one amendment 2014/236/2 dated as of 18 May 2016), and the data were included after written consent from the parents for the acquisition of images and retrospective data analysis.

Informed Consent Statement: Informed parental consent was obtained from all subjects involved in the study.

Data Availability Statement: The original dataset is unavailable due to privacy and ethical restrictions; however, the data analysed and generated are available in the main article and appendices. Further details are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1. BET2 Parameters

To define the parameters for the brain extraction (BE) of neonatal studies, BET2 from FSL version 6.0.6.4 [8,14,44] was tested with modified options from the BET2 program:

- Fractional intensity threshold (<f> parameter varied from 0 to 1 in steps of 0.05; 0.50 is the default value), where smaller values result in larger brain outline estimates, and higher values result in smaller brain outline estimates.
- Vertical gradient in fractional intensity threshold (<g> parameter varied from −1 to 1 in steps of 0.05; 0 is the default value), where positive values result in a larger brain outline at the bottom and a smaller one at the top of the scale, and negative values result in the opposite.
- Robust brain centre estimation (<R>), iterates BET2 several times with the purpose of improving the BE, when the input data contains a considerable amount of non-brain matter.
• Eye and optic nerve cleanup (<S>) attempts to clean up the residual eye and optic nerve voxels that are remaining after running standard BE using BET2.
• Bias field and neck cleanup (<B>) attempts to reduce image bias and residual neck voxels after running standard BE using BET2.

Only one of the options (<R>, <S>, and <B>) can be selected for each BET2 run, together with one represented by different color lines.

The extracted brain images were assessed for quality (i.e., an accurate brain outline, excluding voxels from the eyes, optic nerve, neck, arms, and hands in the neonatal MRI) and according to the dice coefficient (DC) score results, after comparing the results to those with manual segmentation.

After testing all the options, the options (<f> from 0.50 to 0.60 and <g> from −0.10 to 0.20) that provided higher DC scores were pre-selected to perform a further analysis (Figure A1). The DC scores were tested for general differences with the Friedman test ($\chi^2 (20) = 376.364, p < 0.001$). For post hoc comparisons, Bonferroni-adjusted significance tests were used for pairwise comparisons.

The parameters <f> 0.6 and <g> 0.1 were used for all the neonates’ MRI images, as they provided the best DC score (0.936 ± 0.006) and showed no statistically significant differences when compared with other options that also provided high DC scores (>0.931). Including the option <R> also provided good results, without statistically significant differences (paired samples t-test: $t(9) = 1.899; p = 0.071$), but since this option generally runs slower (4.87 ± 0.37 s) compared with the standard (2.20 ± 0.15 s), it was not considered further.

Appendix A.2. SynthStrip Parameters

The BE tool from SynthStrip [26] allows for the fine-tuning of the mask boundary distance from the brain (<b>, 1 mm is the default value). This option was varied from −2 to 3 mm in steps of 1 mm. The extracted brain images were compared with manual segmentation through DC metrics.

The DC were tested for general differences with a repeated measures ANOVA ($F(1,066, 22.396) = 50.975, p < 0.001$, partial $\eta^2 = 0.708$). For post hoc comparisons, Bonferroni-adjusted significance tests were used for pairwise comparisons.

The mask border threshold of 0 and 1 mm provided the highest DC (0.958 ± 0.005 and 0.956 ± 0.008, respectively) and there were no statistically significant differences between them ($p = 1.000$), so in order to simplify processes, the default value (1 mm) was applied to all the neonates’ MRI scans.

All statistical analyses were performed considering a significance value of 5% and using the IBM® SPSS® v27 software.

Figure A1. Dice coefficients with varying BET2 options per subject ($n = 22$ premature neonates, each one represented by different color lines).

The extracted brain images were assessed for quality (i.e., an accurate brain outline, excluding voxels from the eyes, optic nerve, neck, arms, and hands in the neonatal MRI) and according to the dice coefficient (DC) score results, after comparing the results to those with manual segmentation.

After testing all the options, the options (<f> from 0.50 to 0.60 and <g> from −0.10 to 0.20) that provided higher DC scores were pre-selected to perform a further analysis (Figure A1). The DC scores were tested for general differences with the Friedman test ($\chi^2 (20) = 376.364, p < 0.001$). For post hoc comparisons, Bonferroni-adjusted significance tests were used for pairwise comparisons.

The parameters <f> 0.6 and <g> 0.1 were used for all the neonates’ MRI images, as they provided the best DC score (0.936 ± 0.006) and showed no statistically significant differences when compared with other options that also provided high DC scores (>0.931). Including the option <R> also provided good results, without statistically significant differences (paired samples t-test: $t(9) = 1.899; p = 0.071$), but since this option generally runs slower (4.87 ± 0.37 s) compared with the standard (2.20 ± 0.15 s), it was not considered further.

Appendix A.2. SynthStrip Parameters

The BE tool from SynthStrip [26] allows for the fine-tuning of the mask boundary distance from the brain (<b>, 1 mm is the default value). This option was varied from −2 to 3 mm in steps of 1 mm. The extracted brain images were compared with manual segmentation through DC metrics.

The DC were tested for general differences with a repeated measures ANOVA ($F(1,066, 22.396) = 50.975, p < 0.001$, partial $\eta^2 = 0.708$). For post hoc comparisons, Bonferroni-adjusted significance tests were used for pairwise comparisons.

The mask border threshold of 0 and 1 mm provided the highest DC (0.958 ± 0.005 and 0.956 ± 0.008, respectively) and there were no statistically significant differences between them ($p = 1.000$), so in order to simplify processes, the default value (1 mm) was applied to all the neonates’ MRI scans.

All statistical analyses were performed considering a significance value of 5% and using the IBM® SPSS® v27 software.
Appendix B

Table A1. List of segmentation metrics with respective equations and descriptions used for assessment of brain extraction methods.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Equations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voxel-overlap-based metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dice Coefficient (DC) [69,70,92-94]</td>
<td>( DC = \frac{2</td>
<td>VAut ∩ VMan</td>
</tr>
<tr>
<td>Jaccard Coefficient (Jc) [69,70,92-94]</td>
<td>( Jc = \frac{</td>
<td>VAut ∩ VMan</td>
</tr>
<tr>
<td>Precision (Pr) [69,70,94]</td>
<td>( Pr = \frac{</td>
<td>VAut ∩ VMan</td>
</tr>
<tr>
<td>Sensitivity (Se) [69,70,94]</td>
<td>( Se = \frac{</td>
<td>VAut ∩ VMan</td>
</tr>
<tr>
<td>Specificity (Sp) [69,70,94]</td>
<td>( Sp = \frac{TN}{TN + FP} )</td>
<td>Sp is also called true negative rate (TNR). It is defined as the fraction of pixels correctly labelled as &quot;background&quot; (i.e., non-brain tissue).</td>
</tr>
<tr>
<td>False Positive Rate (FPR) [69,70,94]</td>
<td>( FPR = \frac{FP}{FP + TN} )</td>
<td>FPR is also called fallout. It can be considered an estimation of over-segmentation, indicating how many pixels identified as brain, by the automatic method, are outside of the manual mask.</td>
</tr>
<tr>
<td>False Negative Rate (FNR) [69,70,94]</td>
<td>( FNR = \frac{FN}{FN + TP} )</td>
<td>FNR is also called miss rate. It can be considered an estimation of under-segmentation, indicating the falsely segmented pixels over the total number of pixels in manual segmentation.</td>
</tr>
<tr>
<td>Volume Similarity (VS) [70,96]</td>
<td>( VS = \frac{</td>
<td>VAut ∩ VMan</td>
</tr>
<tr>
<td>Root Mean Square Error (RMSE) [67]</td>
<td>( RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Vaут_i - Vман_i)^2} )</td>
<td>RMSE measures the mean (ML) deviations between the automatic and manual segmentations. It is in the square root of the mean error. A smaller RMSE value means better volume agreement between methods.</td>
</tr>
<tr>
<td>Coefficient of Variation of the Root Mean Square Error (CVRMSE) [47]</td>
<td>( CVRMSE = \frac{RMSE}{\text{median} \ V满满} \times 100 )</td>
<td>CVRMSE is also called normalized RMSE or percent RMS. Normalizing the RMSE by the mean value of the manual volume provides the coefficient of variation (%) of the automated volumes relative to the manual volumes.</td>
</tr>
<tr>
<td>Hausdorff Distance (HD) [69,70,96]</td>
<td>( HD(S_{Aut}, S_{Man}) = \max \left( \frac{\sqrt{\sum_{i=1}^{N} (d_i)^2}}{N}, \frac{\sqrt{\sum_{i=1}^{N} (max_{man} - max_{Aut})^2}}{N} \right) )</td>
<td>HD measures the spatial consistency of the overlap between the automatic and manual segmentations by measuring the maximum surface-to-surface distance between the two brain masks. A smaller HD indicates a higher similarity between brain masks.</td>
</tr>
<tr>
<td>Hausdorff Distance 95% Percentile (HD95) [69,70,96]</td>
<td>( HD95(S_{Aut}, S_{Man}) = \max_{man} \left( \frac{\sqrt{\sum_{i=1}^{N} (d_i)^2}}{N}, \frac{\sqrt{\sum_{i=1}^{N} (max_{man} - max_{Aut})^2}}{N} \right) )</td>
<td>HD95 is similar to the maximum HD but considers only the 95% percentiles of the distances in order to overcome the impact of outliers.</td>
</tr>
<tr>
<td>Mean Surface Distance (MSD) [96]</td>
<td>( MSD = \frac{1}{</td>
<td>VAut</td>
</tr>
<tr>
<td>Median Surface Distance (MSD95) [96]</td>
<td>( MSD = \text{median}(d(\text{VAut}_i, \text{VMan}), d(\text{VMan}_i, \text{VAut})) )</td>
<td>MSD95 represents the median of all of the distances from pixels on the boundary of the automatic segmentation to the boundary of the manual segmentation, and vice versa.</td>
</tr>
<tr>
<td>Surface-distance-based metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation Surface Distance (STDSD) [69,96]</td>
<td>( \sqrt{\frac{1}{</td>
<td>VAut</td>
</tr>
</tbody>
</table>

M and A stand for the number of pixels inside the brain mask obtained from manual and automatic segmentations, respectively. \( V_{Man} \) and \( V_{Aut} \) stand for the volume obtained from manual and automatic segmentations, respectively. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) reflect the number of pixels in the automatic segmentation that are classified correctly (true) or incorrectly (false) with respect to the manual segmentation.
Appendix C

Figure A2. Cont.
Figure A2. Comparison of segmentation metrics (a) DC, (b) JC, (c) Pr, (d) Se, (e) Sp, (f) FPR, (g) FNR, (h) VS, (i) HD, (j) HD95, (k) MSD, (l) MDSD) between the four automated brain extraction methods using violin plots (with box plots inside and data points on the side).

References


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.