Review

Deep Learning for MRI-based Acute and Subacute Ischaemic Stroke Lesion Segmentation—A Systematic Review, Meta-Analysis, and Pilot Evaluation of Key Results

Makram Baaklini * and Maria Valdés Hernández

Posted Date: 30 January 2025

doi: 10.20944/preprints202501.2233.v1

Keywords: Acute Ischemic Stroke; MRI; Deep Learning; Neuroimaging; Attention

Review

Deep Learning for MRI-based Acute and Subacute Ischaemic Stroke Lesion Segmentation—A Systematic Review, Meta-Analysis, and Pilot Evaluation of Key Results

Makram Baaklini [1] and Maria del C. Valdés Hernández [2,*]

[1] Edinburgh Imaging Academy, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, UK

[2] Department of Neuroimaging Sciences, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

* Correspondence: Chancellor's Building, 49 Little France Crescent, Edinburgh EH16 4SB, UK. Telephone: +44-131-242 9422, e-mail: M.Valdes-Hernan@ed.ac.uk

Abstract: Background: Segmentation of ischaemic stroke lesions from magnetic resonance images (MRI) remains a challenging task mainly due to the confounding appearance of these lesions with other pathologies, and variations in their presentation depending on the lesion stage (i.e., hyper-acute, acute, sub-acute and chronic). Works on the theme have been reviewed, but none of the reviews have addressed the seminal question on what would be the optimal architecture to address this challenge. We systematically reviewed the literature (2015-2023) for deep learning algorithms that segment acute and/or sub-acute stroke lesions on brain MRI seeking to address this question, meta-analysed the data extracted, and evaluated the results. **Methods and Materials:** Our review, registered in PROSPERO (ID: CRD42023481551), involved a systematic search from January 2015 to December 2023 in the following databases: IEE Explore, MEDLINE, ScienceDirect, Web of Science, PubMed, Springer, and OpenReview.net. We extracted sample characteristics, stroke stage, imaging protocols, and algorithms, and meta-analysed the data extracted. We assessed the risk of bias using NIH's study quality assessment tool, and finally, evaluated our results using data from the ISLES-2015-SISS dataset. **Results:** From 1485 papers, 39 were ultimately retained. 13/39 studies incorporated attention mechanisms in their architecture, and 37/39 studies used the Dice Similarity Coefficient to assess algorithm performance. The generalisability of the algorithms reviewed was generally below par. In our pilot analysis, the UResNet50 configuration, which was developed based on the most comprehensive architectural components identified from the reviewed studies, demonstrated a better segmentation performance than the attention-based AG-UResNet50. **Conclusions:** We found no evidence that favours using attention mechanisms in deep learning architectures for acute stroke lesion segmentation on MRI data, and the use of a U-Net configuration with residual connections seems to be the most appropriate configuration for this task.

Keywords: Acute Ischemic Stroke; MRI; Deep Learning; Neuroimaging; Attention Mechanisms

1. Introduction

Stroke remains a leading cause of mortality and long-term disability worldwide [1], placing a substantial burden on healthcare systems and societies [2]. The majority of strokes are ischaemic [3]. They can occur in different locations and are largely heterogeneous in appearance [3]. After stroke onset, the progression of ischaemic injury continues for minutes-to-days, depending on brain region vulnerability, cellular constituents, and residual perfusion levels [4]. Surrounding the ischaemic core, or irreversibly damaged tissue, appears a region that is functionally impaired, but potentially

 doi:10.20944/preprints202501.2233.v1

salvageable, known as ischaemic penumbra [5]. Accurate diagnosis during acute-to-subacute stages allows for interventions (e.g., thrombolytic drugs, surgery) that may potentially salvage the penumbral area.

Magnetic resonance imaging (MRI) technology has not only enabled the non-invasive investigation of human brain features, but also of ischaemic injuries, thanks to the high dimensionality and particularly low signal-to-noise ratio found in MR images. Segmentation of the infarcted regions in these images, as well as the normal tissues, has been important to advance stroke research and, ultimately, patient outcome. Since manual segmentation methods are time-consuming and subject to inter-rater variability, there has been a growing interest, since 2015 [6], in applying deep learning (DL) techniques to automate stroke lesion segmentation tasks and enhance their accuracy. DL methods can automatically extract intricate spatial and textural features within MR images, while requiring low-to-moderate subject matter expertise. DL also addresses long-dated machine learning-related challenges, such as discerning patterns in high-dimensional data, such as imaging data.

Not surprisingly, several methods have been proposed to automatically assess ischaemic lesions from MRI using DL. These have been analysed previously (Figure 1), but the data that pertains to segmentation of ischaemic stroke lesions have not been systematically reviewed, neither have been meta-analysed, nor their outcomes have been independently evaluated. We systematically review the literature from 2015 to 2023 to investigate the accuracy and generalisability of the proposed DL methods in acute-to-subacute stroke lesion segmentation on MRI, focusing on details of DL architectures and attention mechanisms, seeking to answer the following question: What would be the optimal deep learning model architecture for acute and subacute ischaemic stroke lesion segmentation on brain MRI? After meta-analysing the relevant data extracted from the sources reviewed, we conducted a pilot analysis to evaluate as many of the elements identified in the review as possible.

Reviews*	Scope												Papers overlap ***
	Systematic review	In-depth MA**	Pilot analysis	CT	MRI	DL	ML or pure statistical	Stroke lesion segmentation	WMH/tissue/tumor segmentation	Ischemic stroke	Chronic or hyper-acute stroke	Hemorrhagic stroke	
THIS PAPER	✓	✓	✓		✓	✓		✓	~	✓			N/A
Zhang et al., 2022				✓	✓	✓	✓	✓	✓	✓	✓	~	**11**
Abbasi et al., 2023	✓			✓	✓	✓	~	✓	~	✓	✓	~	**7**
Inamdar et al., 2021	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	**7**
Karthik et al., 2020	✓			✓	✓	✓		✓	~	✓	✓	✓	**7**
Malik et al., 2024				✓	✓	✓		✓		✓	✓		**6**
Liu et al., 2020				✓	✓	✓		✓	✓	✓	✓		**3**
Offersen et al., 2023	✓	~			✓	✓	✓	✓		✓	✓	✓	**2**
Subudhi et al., 2022	✓				✓	✓	✓	✓		✓	✓		**1**
Ciu et al., 2022				✓	✓	✓		✓		✓	✓		**1**
Balakrishnan et al., 2021	✓	~			✓	✓	✓		✓	~			**1**
Lee et al., 2017					✓	✓	✓	✓	~	✓	✓	~	**0**
Karthik et al., 2018					✓	✓	✓	✓		✓	~	✓	**0**
Mainali et al., 2021					✓	✓	✓	✓	✓	✓	✓	✓	**0**
Litjens et al., 2017	✓				✓	✓		✓	✓	✓	✓	✓	**0**
Abang et al., 2020					✓		✓	✓		✓	✓		**0**
Akkus et al., 2017					✓	✓			✓	✓			**0**

In scope	✓
Not in scope	
Partly in scope	~

* Reviews are ordered from "most similar" to "least similar" to our proposed review
** In-depth meta-analysis (MA) = Additional subgroup analysis, sensitivity analysis, meta-regression, and publication bias
*** Papers overlap = Count of papers included in each review, which were also included in our present review

Figure 1. Summary of the scope of the review articles published from 2017 until 2023 that cover similar topics as the present review, and have contributing sources that partially overlap with the ones analysed here.

2. Background

2.1. Deep Learning Architectures

Convolutional neural networks (CNNs) are useful architectures for processing data with grid-like topology (e.g., 2D/3D grid of pixels/voxels) [7]. They employ convolution blocks to produce

 doi:10.20944/preprints202501.2233.v1

"feature maps" through the use of sparse inter-layer interactions, with kernels smaller in size than the input [8]. A standard convolutional block in a CNN (Supplementary Figure S1a) consists of a linear convolution operation on a kernel, which produces a feature map that is passed through an activation function to introduce non-linearity and enable the network to learn more complex relationships in the data [9], before it gets down-sampled by a pooling operation.

CNNs are widely used in medical image segmentation [10], with an architecture that typically ends with fully-connected layer(s) responsible for doing the predictions (e.g., pixel/tissue classification). Predictions are connected to a cost or loss function which measures their discrepancy with ground-truth data. Network parameters are then optimized through backpropagation, by minimizing the loss function until convergence, often aided by regularization methods [9]. However, (i) they produce feature maps with lower spatial dimensions than the input image, and (ii) they classify individual pixels using patches extracted around each pixel, and those often overlap significantly, which in turn creates redundancy in convolution operations. Fully Convolutional Networks (FCNs) address both drawbacks (i) by replacing CNN's fully-connected layer(s) with "up-sampling convolutions" that output images of the same size as the input, and (ii) by generating likelihood maps instead of pixel-by-pixel predictions. However, the FCN's output maps are of particularly low resolution [6].

U-Net architecture was first used for image segmentation in 2015 [11], and it has since achieved overwhelming success. It uses a symmetric encoder-decoder structure based on convolutional blocks, where down-sampling (encoder) operations compress images and up-sampling (decoder) operations restore them, until they reach the input image's original size [12], as opposed to FCNs. U-Nets also introduce skip connections that connect encoder-decoder layers of equal depth, hence allowing them to train with limited data while avoiding the vanishing gradient problem [13].

The ResNet architecture was published shortly after U-Net [14], to further tackle the vanishing gradient problem, also using skip connections. A standard ResNet block (Supplementary Figure S1b) consists of an "identity path" (green arrow in the figure) that can bypass the "residual path", thus giving the network the option to simply copy activations to the next layer and preserve information when learned features do not require more depth. Skip connections also tackle the degradation issue, where adding layers leads to higher training error since accuracy gets "saturated" as the network keeps learning the data [15]. ResNets can improve model convergence speed [16], but since most residual blocks only slightly change the input signal, they produce a large amount of redundant features [17]. This is where DenseNets help.

The first DenseNet architecture was published shortly after ResNet [18]. It employs dense connections interconnecting all layers in order to maximize information and gradient propagation [13]. A standard Dense block is represented in Supplementary Figure S1c. Original inputs and activations from previous layers are both kept at each block, hence preserving the global state, while encouraging feature reuse with less network parameters [12]. Reusing features across layers also allows DenseNets to tackle the vanishing gradient problem [19].

2.2. Attention Mechanisms

When our eyes focus on a certain object, groups of filters within our visual perception system are used to create a blurring effect so that the object of interest is in focus, and the rest is blurred [20]. Attention mechanisms attempt to achieve the same "blurring effect" but for machine-based image processing. Attention can capture the large receptive field and retrieve underlying contextual details by modelling the relationships between local and global features [21]. In this work, we categorize attention mechanisms as "spatial", "channel", or "hybrid".

"Spatial attention" (Supplementary Figure S2a) is responsible for generating masks that enhance the features that define a specified object (e.g., lesion) on a given feature map, therefore enhancing the input to subsequent layers of a network [20]. Examples of spatial attention methods include attention gates, i.e., computational blocks to implement "attention" as described above; self-attention, which operates solely on input sequences, thus enabling a model to further exploit spatial

 doi:10.20944/preprints202501.2233.v1

relationships within input scans [22]; and cross-attention (e.g., Gomez et al. [23]), which enables the network to simultaneously process encoder and decoder features, in order to pass the most aligned encoder features with respect to decoder features of same depth, and therefore decrease noisy signals in skip connections [22].

"Channel attention" (Supplementary Figure S2b) refers to the process of assigning a weight to each feature map or channel, emphasizing those that contribute most significantly to the learning [20]. Conversely, spatial attention assigns weights to pixels. Each map specializes in detecting specific features (e.g., horizontal edges, brain anatomy). Examples of channel attention methods include squeeze-and-excitation blocks [24], which were used by Woo et al. [25] and Lee A. et al. [26]).

In summary, channel attention focuses on the importance of different feature maps, while spatial attention focuses on the importance of specific regions within a feature map.

"Hybrid attention" combines spatial and channel attention. Examples include dual attention gates, which combine spatial and channel attention gates (sAG+cAG) [27]; and multi-head attention, which uses parallel processing by applying attention across multiple "heads" simultaneously, where each head may be configured to implement any channel or spatial attention operation [22].

3. Materials & Methods

3.1. Protocol Registration

We registered this systematic review protocol with the International Prospective Register of Systematic Reviews (PROSPERO), registration number: CRD42023481551 (November 2023). We conducted our review following the PRISMA guidelines [28,29].

3.2. Search Strategy

We conducted a literature search (January 2015 – December 2023) for papers published in IEEE Explore, MEDLINE, ScienceDirect, Web of Science, PubMed, Springer, and OpenReview.net. We identified keywords by expanding five subject components: accuracy, acute ischaemic stroke, deep learning, lesion segmentation, and MRI.

We also did citation tracking of reviewed articles, and hand-searching of the two journals "Stroke" and "NeuroImage: Clinical" (Recall:100%). Two reviewers (M.B. and M.C.V.H.) conducted the main search, paper selection, and data extraction, and discrepancies were resolved by discussion. The full search strategy is provided in Supplementary Data A.

3.3. Eligibility Criteria

Table 1 summarizes the selection criteria, justifying the basis for inclusion and exclusion of the different articles found during the search. Briefly, studies were included if presented (a) deep-learning algorithm(s)/architecture(s) for segmenting ischaemic stroke lesions in acute and subacute phases in humans, from MRI, and were peer-reviewed and indexed in any of the databases searched. Studies were excluded otherwise.

Table 1. Study selection criteria.

	Included	**Excluded**	**Rationale**
Stroke types	Ischaemic	Haemorrhagic	Differences in clinical presentations, lesion appearances, & aetiologies.

 doi:10.20944/preprints202501.2233.v1

Stroke stages	• Acute • Subacute	• Hyperacute (unless in minor proportion in the dataset) • Chronic	Prioritize stages where MRI plays a more prominent role in diagnosis and treatment planning.
Imaging	• All MRI modalities • All scanner types	• All CT modalities • Any other non-MRI modality	MRI allows in-vivo assessment offering better soft tissue contrast & resolution than CT and PET.
Algorithms	• All deep-learning approaches (e.g., supervised, unsupervised) • Algorithms segmenting both: ischaemic core and penumbra	• Non-deep learning algorithms • Algorithms segmenting only WMH or brain tissue/tumours • Algorithms performing semi-automated segmentation (with human interaction) • Algorithms running on simulated/synthetic lesions	DL is the current state-of-the-art computational approach, much better than others at learning complex hierarchical features.
Population	Humans (all ages/sexes).	• Non-human studies (e.g., animal-based) • Human studies using synthetic data	Human-based studies are more clinically relevant. Synthetic data may not fully capture variations and complexities of real clinical stroke lesions.
Publishing	• Peer-reviewed studies • Proceedings of MICCAI, MIDL, and IEEE-led conferences • Publications in English • Publications between 2015-2023	• Pre-prints • Studies not available in any of the searched databases	To only retain the most reliable sources of information while also aiming for a wide readership.
Completeness	Studies with sufficient information to be reproduced.	Studies not reporting segmentation performance scores.	Reproducibility is key in scientific research.

 doi:10.20944/preprints202501.2233.v1

3.4. Data Extraction

For each paper, we extracted the following information: primary outcomes and measures, image acquisition protocol(s), sample characteristics, ground-truth data, data pre-processing, learning approach, model architecture, model training, model hyper-parameters, model validation, external validation, performance results, and generalisability of the proposed approach as per custom calculation. To cross-check data entry, a reviewer (M.C.V.H.) performed double extraction independently and blind to prior extraction results.

3.5. Data Analysis

We analysed the extracted results using custom-built scripts in python. We calculated fixed-effects and random-effects as part of a whole group analysis. For these analyses we used the reported dice similarity coefficients (DSC) and their 95% confidence intervals (CI) to estimate the effect size. For the effect estimates we used the weighted average of the reported mean DSC. We further divided the studies in two groups: (i) studies using attention mechanisms, and (ii) studies not using attention mechanisms and repeated the analyses for each group. We also conducted a sensitivity analysis using the precision metrics (instead of the DSC) to estimate the effect size. Lastly, we conducted a meta-regression analysis to assess whether there is statistically significant relationship between the presence of attention mechanisms and the likelihood of high mean DSC across studies. We further used the DSC and the standard errors for generating a funnel plot, followed by the Egger's test, to assess possible bias in the meta-analysis.

3.6. Publication Quality Analysis

We assessed the sources selected following the NIH's Study Quality Assessment Tool (https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools).

3.7. Pilot Analysis

We conducted a pilot analysis leveraging the findings from our literature analysis in an independent and publicly available sample. The specific aims of this pilot were two-fold:

- Proposing an architecture that leverages the findings of our systematic review in terms of best development practices:
  - Use 2D model with image-wise training
  - Increase network depth while leveraging the power of skip connections by combining U-Net and ResNet
- Conducting multiple experiments (24 in total) assessing segmentation performance in different scenarios:
  - With versus without attention mechanisms
  - Using a compound loss function versus a region-based loss function
  - Using input images of a single modality (DWI) versus input images of multiple modalities

3.7.1. Dataset

We used the ISLES-2015-SISS dataset, published by the MICCAI 2015 conference [30]. It consists of brain MRI from 28 subacute stroke cases to use for model training. For each case, a set of five MRI sequences are provided: T1-weighted (T1-WI), T2-weighted (T2-WI), diffusion-weighted (DWI), and fluid-attenuated inversion recovery (FLAIR), along with the corresponding ground-truth masks. The data were already anonymized by removing patient information from files and facial bone structure from images.

3.7.2. Data Pre-Processing

 doi:10.20944/preprints202501.2233.v1

The following data pre-processing steps were conducted:

- Intensity-based normalization using Min-Max scaling
- Intensity-based skull-stripping using BET2 (performed by challenge organizers)
- Rigid co-registration to the FLAIR sequences (performed by challenge organizers)

3.7.3. Segmentation Architecture, Model Training and Evaluation

We implemented the DL architecture, AG-UResNet50, inspired by multiple papers [31–37], especially Guerrero et al.'s UResNet [34], Jin et al.'s RA-UNet [36], and Gheibi et al.'s CNN-Res [37]. AG-UResNet50 is a five-level end-to-end U-Net (Figure 2), with a ResNet50 replacing its encoder path [38]. Using U-Net in combination with ResNet50 allows us to leverage the power of skip connections further [39], and make the network deeper. This makes it easier for the gradient to flow from output layers back to input during back-propagation, while handling the vanishing gradient problem. Zhang et al. [40] identified ResNet as an architecture that can improve segmentation of small lesions. Max-pooling was used for down-sampling the first set of feature maps produced by the model, since it can extract extreme features (e.g., lesion edges) well. Convolution blocks with stride two were used for remaining down-sampling operations, in order to better retain image details [13]. On the decoder side, we simply used the U-Net's deconvolution blocks, but with Leaky ReLU activation instead of ReLU, in view of its better results in medical image analysis [41], as also demonstrated by Karthik et al. [42]. We kept the up-sampling interpolation algorithm, which basically inserts new elements between pixels in the image matrix. Feature maps from the encoder are combined with those from the decoder in the same depth using concatenation. "Attention concatenation", which was used here, works by incorporating attention gates (AGs) in skip connections [43], as seen in Karthik R. et al. [44], Nazari-Farsani et al. [45], and Yu et al. [46]. An AG takes two input vectors that are added element-wise (Figure 3), resulting in aligned weights becoming larger and unaligned weights smaller. The output vector then goes through ReLU activation, 1x1 convolution, and sigmoid activation to produce the attention coefficients/weights. Coefficients are then up-sampled to the original dimensions of the input vector using trilinear interpolation, before being multiplied element-wise. The final output is passed along in the skip connection.

During training, we used a compound loss function mixing Binary Cross-Entropy (BCE) and Dice loss. BCE loss computed the gradient based on the difference in probability distribution of each pixel in the predicted versus real sample [47], while Dice loss directly computed the gradient using the Dice score of predicted versus real samples [18]. From a regularization standpoint, we used pixel dropout, learning rate adjustment and data augmentation methods, while for optimization, we used Adam function and batch normalization. From a training infrastructure standpoint, the model was developed, trained and tested on Azure Databricks (python:Torch), using one sizeable driver: CPU:16 cores; OS:Ubuntu; RAM:56GB; Runtime:13.2ML. We evaluated the model performance using DSC, and used five-fold cross-validation. The full code used for this pilot is available from GitHub (https://github.com/Elpazzu/UoE-Pilot-Analysis/)

 doi:10.20944/preprints202501.2233.v1

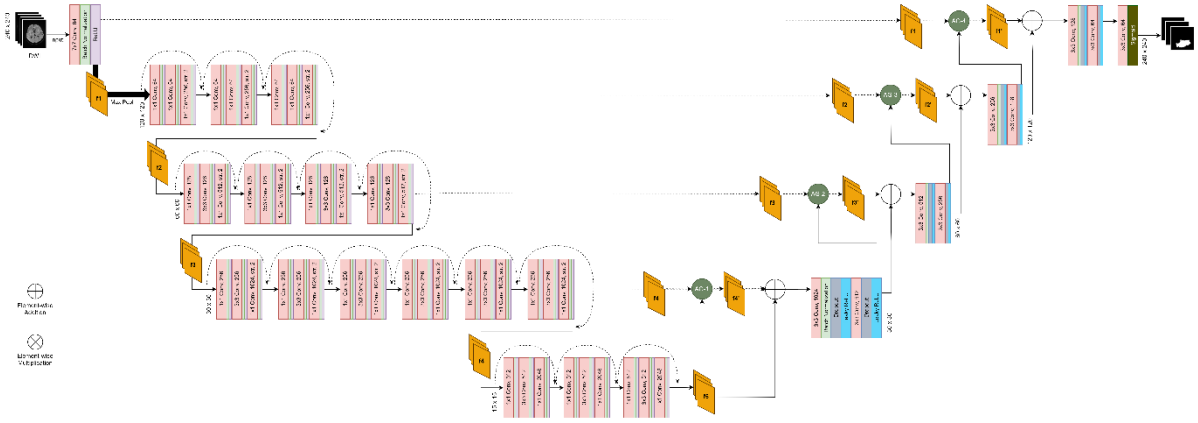

Figure 2. Architecture of AG-UResNet50. Please refer to **Supplementary Data B1** for a zoomed in view of this diagram, with a more detailed description of its architectural components.

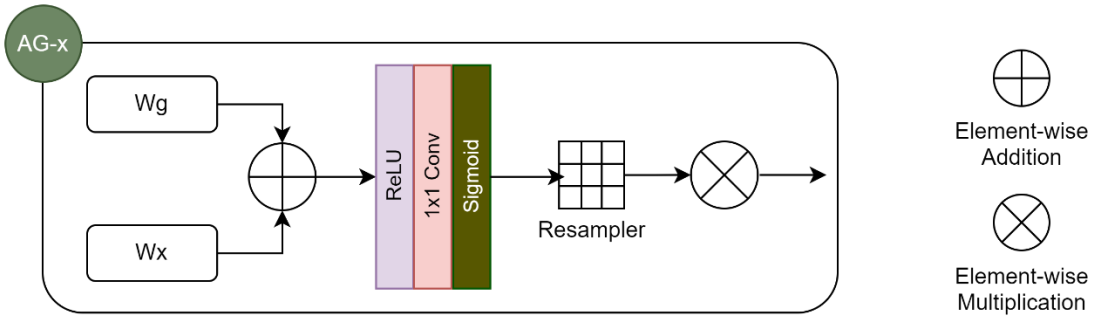

Figure 3. Architecture of an Attention Gate (AG), as used in our pilot analysis.

4. Results

4.1. Search Results

The search yielded 1485 papers, of which 39 were ultimately retained (**Figure 4**).

 doi:10.20944/preprints202501.2233.v1

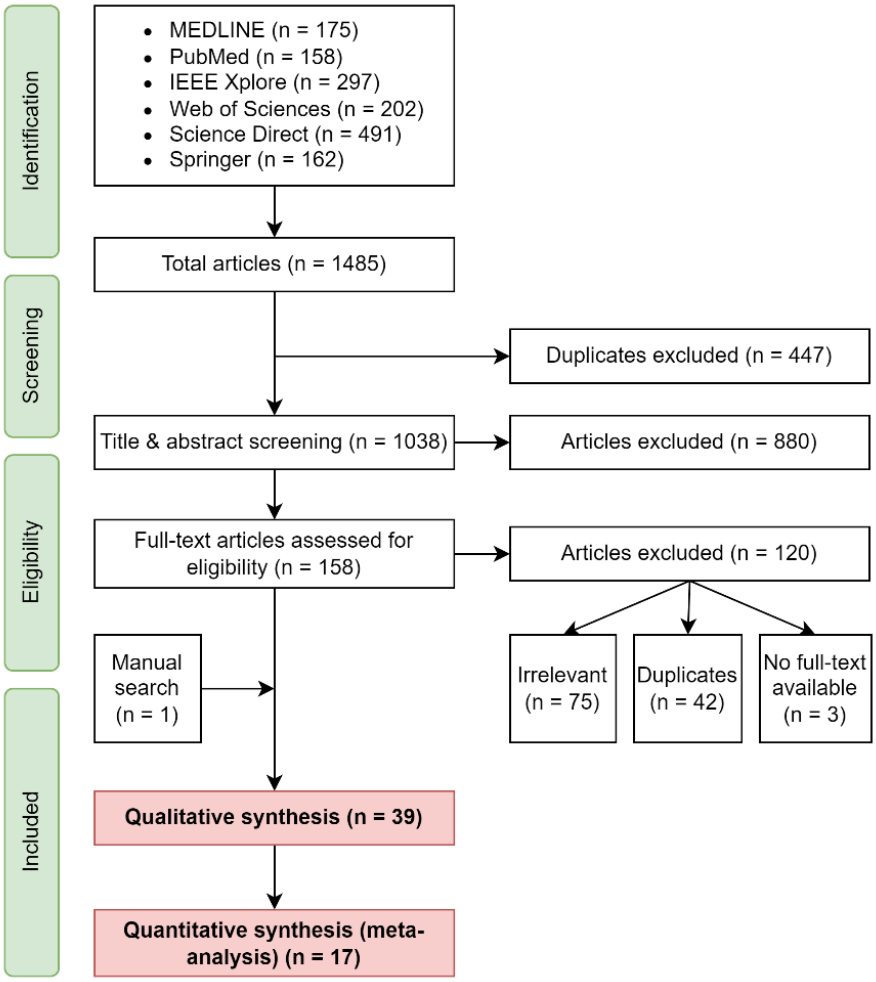

Figure 4. Flow chart of the identification, screening, and paper selection process.

All papers had segmentation as primary outcome. Fewer had prognosis (6 studies) or functional (3 studies) outcomes. Prognosis studies were either trying to predict tissue fate or lesion volume (e.g., Wong et al. [48], Wei et al. [49]). Functional studies mostly tried to predict the modified ranking scale score (mRS). Only one paper explicitly had diagnosis as primary outcome, but practically, segmentation and diagnosis are tightly linked, since by segmenting lesion pixels, the algorithm is effectively helping physicians with the diagnosis.

4.2. Sample Characteristics

As Table 2 shows, patients were all adults of 18 years old and above, and males were generally slightly over-represented (58% on average), except in few studies where the opposite was true (e.g., Moon et al. [50]). From a stroke severity standpoint, reported mean NIHSS [51] were always on the "minor" or "moderate" ranges (7 studies). Although both subacute and acute stroke stages were in scope, most studies (23/39) included exclusively acute ischaemic stroke cases. Reported patient mean "time-since-stroke" (TSS) were also exclusively in the acute interval, with 2 studies actually very close to the hyperacute-acute limit. Only four papers used sample sizes above 500 (Mean 252.2), and samples were most often collected from only one center (28 studies versus 14 leveraging multiple centers). Supplementary Figure S3 shows a graphical illustration of the sample characteristics.

Table 2. Characteristics of the samples of the studies included in the review.

 doi:10.20944/preprints202501.2233.v1

First author	Sample size	Number of medical centers	Stroke stage	Age range	Gender	Mean NIHSS	Mean Stroke-to-MRI time	Mean lesion volume	Lesion volume ranges
		SC: Single-center; MC: Multi-center	*Acute; Subacute*		*M: Male F: Female*			*(in ml)*	*(in ml)*
Karthik, R. [42]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Gómez, S. [23]	75	**MC**: 2	Acute	[18+]	-	-	-	37.83	[1.6, 160.4]
Olivier, A. [70]	929	**MC**: 6	Acute Subacute	[16–94]	M: 63.7% F: 36.3%	7.6	68.8h	21.84	-
Clèrigues, A. [69]	114	**MC**: 4	Acute Subacute	[18+]	-	-	-	SISS: 17.59 SPES: 133.21	SISS: [1.0, 346.1] SPES: [45.6, 252.2]
Liu, L. [60]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Moon, H. [50]	79	-	Acute	-	M: 44.3% F: 55.7%	9.3	83.8h	-	[0.0, 250]
Zhang, R. [19]	242	**SC**: 1	Acute	[35–90]	M: 60.3% F: 39.7%	-	-	-	-
Wong, K. [48]	875	**SC**: 1	Acute	-	M: 48.9% F: 51.1%	6	-	-	-
Khezrpour, S. [79]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Hu, X. [90]	75	**MC**: 2	Acute	[18+]	-	-	-	37.83	[1.6, 160.4]
Gheibi, Y. [37]	44	**MC**: 2	Acute	-	-	-	-	-	-

 doi:10.20944/preprints202501.2233.v1

Kumar, A. [110]	189	**MC**: 6	Acute Subacute	[18+]	-	-	-	SISS: 17.59 SPES: 133.21 IS17: 37.83	SISS: [1.0, 346.1] SPES: [45.6, 252.2] IS17: [1.6, 160.4]
Liu, L. [16]	79	**MC**: 2	Acute	[18+]	-	-	-	SPES: 133.21 LHC: -	SPES: [45.6, 252.2] LHC: -
Zhao, B. [71]	582	**SC**: 1	Acute	-	-	-	-	-	-
Liu, C. [27]	1849	**SC**: 1	Acute Subacute	[52–73]	M: 52.9% F: 47.1%	3.4	17.7h	3.12	[1.55, 5.33]
Karthik, R. [44]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Liu, L. [95]	114	**MC**: 4	Acute Subacute	[18+]	-	-	-	SISS: 17.59 SPES: 133.21	SISS: [1.0, 346.1] SPES: [45.6, 252.2]
Aboudi, F. [59]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Pinto, A. [97]	75	**MC**: 2	Acute	[18+]	-	-	-	37.83	[1.6, 160.4]
Choi, Y. [96]	54	**MC**: 2	Acute	[18+]	-	-	-	37.83	[1.6, 160.4]
Kim, Y. [63]	296	**SC**: 1	Acute	[58–79]	M: 61.3% F: 38.7%	2.3	12.7h	12.19	[0.0, 279.4]
Woo, I. [25]	429	**SC**: 1	Acute	[24–98]	M: 62.3% F: 37.7%	-	21.4h	-	-
Lee, A. [26]	429	**SC**: 1	Acute[24–98	[24–98]	M: 62.3% F: 37.7%	-	21.4h	27.44	[0.3, 227.6]

 doi:10.20944/preprints202501.2233.v1

Lee, S. [81]	472	**SC**: 1	Acute	[19+]	M: 63.3% F: 36.7%	3	4.9h	3.62	[0.52, 71.8]
Karthik, R. [101]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Zhang, L. [99]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Ou, Y. [103]	99	**SC**: 1	Acute	-	-	-	-	-	-
Vupputuri, A. [102]	189	**MC**: 6	Acute Subacute	[18+]	-	-	-	SISS: 17.59 SPES: 133.21 IS17: 37.83	SISS: [1.0, 346.1] SPES: [45.6, 252.2] IS17: [1.6, 160.4]
Abdmouleh, N. [64]	64	**MC**: 3	Subacute	[18+]	-	-	-	17.59	[1.0, 346.1]
Duan, W. [98]	120	**SC**: 1	Acute	-	-	-	-	-	-
Lucas, C. [100]	75	**MC**: 2	Acute	[18+]	-	-	-	37.83	[1.6, 160.4]
Nazari-Farsani, S. [45]	445	**MC**: 6+	Acute	-	M: 50% F: 50%	13	6.2h	50	[15, 123]
Wei, Y. [49]	216	**SC**: 1	Acute	-	M: 69.7% F: 30.3%	-	-	-	-
Li, C. [72]	60	**SC**: 1	Acute	[49–88]	-	-	-	-	-
Liu, Z. [67]	212	**SC**: 1	Acute Subacute	-	M: 62% F: 38%	-	-	-	-
Cornelio, L. [58]	75	**MC**: 2	Acute	[18+]	-	-	-	37.83	[1.6, 160.4]
Yu, Y. [46]	182	**MC**: 6+	Acute	-	M: 46.7% F: 53.3%	15	-	54	[16, 117]
Wu, Z. [87]	400	**MC**: 3	Subacute	[18+]	-	-	-	27.94	[0.0575, 340.28]

 doi:10.20944/preprints202501.2233.v1

Guerrero, R. [34]	250	**SC**: 1	Acute	-	-	-	-	-	-

4.3. Imaging Acquisition and Manipulation

Table 3 shows the imaging data extracted from the reviewed sources, and Figure 6 plots the correspondence between the dimensions of the images used as input to the reviewed algorithms (i.e., 2D, 2.5D, or 3D) and the spatial resolution and the manipulation of these images during training (i.e., patch-wise or image-wise). Most studies (25/39) used images of high or very high spatial resolution. DWI modality was by far the most used modality (37 studies), followed by FLAIR (17 studies). Also, two-thirds of studies adopted a multi-modal approach. Twenty-seven studies used a 2D-based approach and 10 a 3D-based approach (Table 3). 2D models exclusively used high or very high-resolution images, and 3D models exclusively moderate or low-resolution images, which seems counter intuitive (Figure 6a). 3D models adopted patch-wise training in 8/10 studies (Figure 6b). During the generation of the ground-truth, 31/39 studies reporting mismatch between image sequences; 13 studies leveraged diffusion-perfusion (DWI-PWI) mismatch, and 12 DWI-FLAIR mismatch. The magnetic field of the scanner(s) was 1.5T and 3T in 25 studies, only 3T in nine studies, and only 1.5T in three studies. See pie charts in Supplementary Data B.

Table 3. Imaging acquisition and manipulation in the reviewed studies.

First author	Spatial resolution	Image modalities	Input dimension	Modality mismatch	Magnetic field
	1-Very High (VH); 2-High (H); 3-Moderate (M); 4-Low (L)	*SM: Single-modality; MM: Multi-modality* *Format: Modality-{Parameter}*	*2D; 2.5D; 3D*	*T1-T2; DWI-PWI; DWI-FLAIR; T2-FLAIR; T1-FLAIR*	*1.5T; 3T*
Karthik, R. [42]	2-H	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000}	2D	DWI-FLAIR	3T
Gómez, S. [23]	2-H	**MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	2D	DWI-PWI	1.5T 3T
Olivier, A. [70]	Not reported	**SM**: {DWI-b0, DWI-b1000, DWI-ADC}	3D	None reported	1.5T 3T
Clèrigues, A. [69]	4-L	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000} **MM**: {T1WI, T2WI, DWI-b1000, PWI-CBF, PWI-CBV, PWI-TTP, PWI-Tmax}	3D	DWI-PWI; DWI-FLAIR	1.5T 3T
Liu, L. [60]	2-H	**MM**: {FLAIR, DWI-b1000}	2D	DWI-FLAIR	3T
Moon, H. [50]	1-VH	**MM**: {FLAIR, DWI-b1000}	2D	None reported	1.5T
Zhang, R. [19]	3-M	**SM**: {DWI-b0, DWI-b1000, DWI-ADC}	3D	None reported	1.5T 3T

 doi:10.20944/preprints202501.2233.v1

Wong, K. [48]	Not reported	**SM**: {DWI-b0, DWI-b1000, DWI-eADC}	2D	None reported	1.5T 3T
Khezrpour, S. [79]	2-H	**SM**: {FLAIR}	2D	DWI-FLAIR	3T
Hu, X. [90]	4-L	**MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	3D	DWI-PWI	1.5T 3T
Gheibi, Y. [37]	Not reported	**MM**: {FLAIR, DWI}	2D	None reported	-
Kumar, A. [110]	4-L	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000} **MM**: {T1WI, T2WI, DWI-b1000, PWI-CBF, PWI-CBV, PWI-TTP, PWI-Tmax} **MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	3D	DWI-PWI; DWI-FLAIR	1.5T 3T
Liu, L. [16]	2-H	**MM**: {T1WI, T2WI, DWI-b1000, PWI-CBF, PWI-CBV, PWI-TTP, PWI-Tmax} **MM**: {DWI, T2WI}	2D	DWI-PWI	1.5T 3T
Zhao, B. [71]	2-H	**SM**: {DWI-ADC, DWI-b0, DWI-b1000}	2D	None reported	1.5T 3T
Liu, C. [27]	3-M	**SM**: {DWI-b0, DWI-ADC, DWI-IS}	3D	None reported	1.5T 3T
Karthik, R. [44]	2-H	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000}	2D	DWI-FLAIR	3T
Liu, L. [95]	2-H	**MM**: {FLAIR, DWI-b1000} **MM**: {T2WI, DWI-b1000, PWI-CBF, PWI-CBV, PWI-TTP, PWI-Tmax}	2D	DWI-PWI; DWI-FLAIR	1.5T 3T
Aboudi, F. [59]	2-H	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000}	2D	DWI-FLAIR	3T
Pinto, A. [97]	2-H	**MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	2D	DWI-PWI	1.5T 3T
Choi, Y. [96]	4-L	**MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	3D	DWI-PWI	1.5T 3T
Kim, Y. [63]	1-VH	**SM**: {DWI-b0, DWI-b1000, DWI-ADC}	2D	None reported	1.5T 3T
Woo, I. [25]	1-VH	**SM**: {DWI-b1000, DWI-b0, DWI-ADC}	2D	None reported	1.5T 3T

 doi:10.20944/preprints202501.2233.v1

Lee, A. [26]	1-VH	**SM**: {DWI-b1000, DWI-b0, DWI-ADC}	2D	None reported	1.5T 3T
Lee, S. [81]	3-M	**MM**: {DWI, DWI-ADC, FLAIR, PWI-Tmax, PWI-TTP, Pred(init)}	3D	DWI-PWI	1.5T 3T
Karthik, R. [101]	2-H	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000}	2D	DWI-FLAIR	3T
Zhang, L. [99]	2-H	**SM**: {DWI-b1000}	2D	DWI-FLAIR	3T
Ou, Y. [103]	1-VH	**SM**: {DWI-b1000, DWI-eADC}	2.5D	None reported	1.5T 3T
Vupputuri, A. [102]	2-H	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000} **MM**: {T1WI, T2WI, DWI-b1000, PWI-CBF, PWI-CBV, PWI-TTP, PWI-Tmax} **MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	2D	DWI-PWI; DWI-FLAIR	1.5T 3T
Abdmouleh, N. [64]	2-H	**MM**: {FLAIR, T2WI, T1WI, DWI-b1000}	2D	DWI-FLAIR	3T
Duan, W. [98]	Not reported	**MM**: {T2WI, DWI-b1000, DWI-b0}	3D	None reported	-
Lucas, C. [100]	2-H	**MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	2D	DWI-PWI	1.5T 3T
Nazari-Farsani, S. [45]	Not reported	**SM**: {DWI-b1000, DWI-ADC}	3D	None reported	1.5T 3T
Wei, Y. [49]	1-VH	**SM**: {DWI-b1000}	2D	T1-T2; T2-FLAIR; T1-FLAIR	3T
Li, C. [72]	1-VH	**MM**: {T1WI, T2WI, T2WI-FLAIR, DWI, DWI-ADC}	2D	T2-FLAIR; T1-FLAIR	1.5T
Liu, Z. [67]	Not reported	**MM**: {T2WI, DWI, DWI-ADC}	2D	None reported	1.5T 3T
Cornelio, L. [58]	2-H	**MM**: {DWI-ADC, PWI-rCBF, PWI-rCBV, PWI-MTT, PWI-TTP, PWI-Tmax, Raw 4D PWI}	2D	DWI-PWI	1.5T 3T
Yu, Y. [46]	Not reported	**MM**: {DWI-b1000, DWI-ADC, PWI-Tmax, PWI-MTT, PWI-CBF, PWI-CBV}	2.5D	None reported	1.5T 3T
Wu, Z. [87]	1-VH	**MM**: {DWI-b1000, DWI-ADC, FLAIR}	2D	DWI-FLAIR	1.5T 3T

 doi:10.20944/preprints202501.2233.v1

Guerrero, R. [34]	1-VH	**MM**: {FLAIR, T1WI}	2D	None reported	1.5T

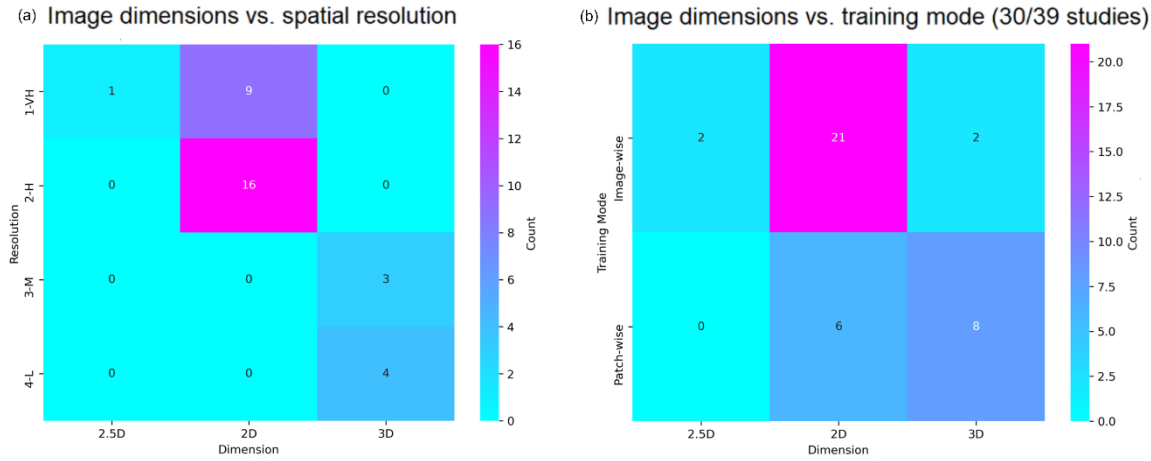

Figure 6. a) Correlation between the dimension and the spatial resolution of input images; **b)** Correlation between the dimension of input images and the adopted model training mode.

4.4. Data Pre-Processing

Eighteen studies used proprietary datasets (Table 4), 24 used one or a combination of ISLES-2015 [30], ISLES-2017 [52] or ISLES-2022 [53], and three used data related to the DEFUSE or iCAS studies [54–56]. In relation to skull-stripping, 37 studies performed an intensity-based approach (using BET2/ITK software), one study used an atlas-based approach (Moon et al. [50] using Kirby/MMRR template), and one study used DL to reduce sensitivity-to-noise [57] (Liu C. et al. [27] using in-house "UNet BrainMask"). Inter-patient image registration onto a standard space (e.g., MINI) and/or intra-patient registration (e.g., registration of different sequences) were performed in 27 studies.

Table 4. Data pre-processing in the reviewed studies.

		Data pre-processing methods				
First author	**Dataset**	**Intensity-based**	**Atlas-based**	**Morphology-based**	**Deformable surface-based**	**Machine learning-based**
	Dataset used for model training	*Data pre-processing techniques used prior to model training*				
Karthik, R. [42]	ISLES2015 SISS	- Normalization - Skull-stripping	-	- Resizing	- Registration	-
Gómez, S. [23]	ISLES2017	- Normalization - Contrast adjustment - Skull-stripping	-	- Resizing - Rescaling	- Registration	-
Olivier, A. [70]	Proprietary	- Normalization	-	- Rescaling - Zero-	-	-

 doi:10.20944/preprints202501.2233.v1

				padding - Cropping		
Clèrigues, A. [69]	ISLES2015 SISS ISLES2015 SPES	- Normalization - Skull-stripping	-	-	- Registration	-
Liu, L. [60]	ISLES2015 SISS	- Skull-stripping	-	-	- Registration	-
Moon, H. [50]	Proprietary	- Normalization	- Skull-stripping	- Zero-padding - Resizing	- Registration	-
Zhang, R. [19]	Proprietary	- Normalization	-	- Zero-padding - Cropping - Resizing	-	-
Wong, K. [48]	Proprietary	- Normalization	-	-	-	-
Khezrpour, S. [79]	ISLES2015 SISS	- Contrast adjustment - RGB to greyscale - Skull-stripping	-	- Cropping - Resizing	- Registration	-
Hu, X. [90]	ISLES2017	- Skull-stripping	-	- Resizing - Cropping	- Registration	-
Gheibi, Y. [37]	Proprietary	-	-	- Zero-padding	- Splitting into 2D	-
Kumar, A. [110]	ISLES2015 SPES ISLES2015 SSIS ISLES2017	- Normalization - Skull-stripping		- Resizing	- Converting to 3D - Registration	- Slice classification
Liu, L. [16]	ISLES2015 SPES Proprietary	- Normalization - Smoothing - Skull-stripping	-	-	- Registration	-
Zhao, B. [71]	Proprietary	- Normalization	-	-	-	-
Liu, C. [27]	Proprietary	- Normalization	-	- Resizing - Rescaling	-	- Slice classification - Skull-stripping
Karthik, R. [44]	ISLES2015 SISS	- Skull-stripping	-	-	- Registration	-

 doi:10.20944/preprints202501.2233.v1

Liu, L. [95]	ISLES2015 SPES ISLES2015 SISS	- Normalization - Skull-stripping	-	- Zero-padding - Cropping - Resizing	- Registration - Splitting into 2D	-
Aboudi, F. [59]	ISLES2015 SISS	- RGB to greyscale - Skull-stripping	-	- Resizing - Rescaling	- Registration	-
Pinto, A. [97]	ISLES2017	- Normalization - Bias field correction - Skull-stripping	-	- Resizing	- Registration	-
Choi, Y. [96]	ISLES2016	- Normalization - Skull-stripping	-	- Resizing - Rescaling	- Registration	-
Kim, Y. [63]	Proprietary	- Normalization	-	- Resizing	-	-
Woo, I. [25]	Proprietary	- Normalization	-	-	- Registration	-
Lee, A. [26]	Proprietary	- Normalization	-	- Resizing	- Registration	-
Lee, S. [81]	Proprietary	-	-	- Resizing - Rescaling	- Registration	-
Karthik, R. [101]	ISLES2015 SISS	- Normalization - Skull-stripping	-	- Cropping - Rescaling	- Registration	- Slice classification
Zhang, L. [99]	ISLES2015 SISS	- Normalization - Skull-stripping	-	- Cropping - Rescaling	- Registration	- Slice classification
Ou, Y. [103]	Proprietary	- Normalization - Skull-stripping	-	- Resizing	-	-
Vupputuri, A. [102]	ISLES2015 SPES ISLES2015 SISS ISLES2017 (IS17)	- RGB to greyscale - Normalization - Skull-stripping	-	-	- Registration	-
Abdmouleh, N. [64]	ISLES2015 SISS	- Normalization - Skull-stripping	-	-	- Registration	-
Duan, W. [98]	Proprietary	- Normalization - Skull-stripping	-	- Resizing	-	-
Lucas, C. [100]	ISLES2017	- Skull-stripping	-	- Rescaling	- Registration	-
Nazari-Farsani, S. [45]	UCLA iCAS DEFUSE DEFUSE-2	- Normalization	-	-	- Registration	-

 doi:10.20944/preprints202501.2233.v1

Wei, Y. [49]	Proprietary	- Skull-stripping	-	- Rescaling	- Registration	-
Li, C. [72]	Proprietary	-	-	- Resizing	-	-
Liu, Z. [67]	Proprietary	- Normalization - Skull-stripping	-	- Cropping - Resizing	- Registration	-
Cornelio, L. [58]	ISLES2017	- RGB to greyscale - Contrast adjustment - Normalization - Skull-stripping	-	- Resizing	- Registration	-
Yu, Y. [46]	iCAS DEFUSE-2	- Normalization	-	-	- Registration	-
Wu, Z. [87]	ISLES2022	- Skull-stripping	-	- Resizing - Rescaling	- Registration	-
Guerrero, R. [34]	Proprietary	- Normalization	-	- Resizing	- Registration	-

4.5. Deep Learning Architectures

Within the 37/39 studies that performed semantic segmentation, 35 studies used U-Net-based models (Figure 7). But none of them used the original U-Net *as-is* [11], with perhaps Cornelio et al. [58] and Aboudi et al. [59] being the closest. ResNet architecture was the second most used (8 studies), while DenseNets were only used in three studies. Data augmentation was the most used regularization method (28 studies), whereas each of dropout, early stopping, weight decay, class weighting, and learning rate adjustment were used in 10-11 studies. More papers used image-wise training (27 studies versus 14 for patch-wise training); 7/8 studies that were dealing with smaller mean lesion volumes (< 40ml) used patch-wise training. In addition, none of the papers performed uncertainty quantification, and 31 algorithms were end-to-end (versus 8 multi-module). Twenty-three studies used Dice loss (Table 5), either mixed with other loss functions (8 studies) or standalone (15 studies). Cross-entropy loss was used in 17 papers, nine times standalone. Focal loss was only used in four papers, and two papers used Liu et al.'s custom-built loss function [16]. Twelve studies used attention: five used hybrid attention, four spatial attention, and three used channel attention (Table 5). Studies deploying ResNet-based architectures did not incorporate attention.

Table 5. Deep learning architectures of the models presented in the studies included (see corresponding summary graphs in the Supplementary Data B).

First author	Architecture (Segmentation type)	Loss function	Attention mechanism / type	Activation functions	Regularization method	Optimization method	Epochs
Karthik, R. [42]	U-Net (Semantic)	Dice	None	Leaky ReLU, ReLU, Softmax	Data augmentation	Adam	120

 doi:10.20944/preprints202501.2233.v1

Gómez, S. [23]	U-Net (Semantic)	Focal	Additive cross-attention / spatial	ReLU, Sigmoid	- Data augmentation - Weight decay - Class weighting	AdamW	600
Olivier, A. [70]	U-Net (Semantic)	Dice	None	Leaky ReLU, Softmax	- Data augmentation - ES on validation loss	Adam	-
Clèrigues, A. [69]	U-Net (Semantic)	Focal	None	PReLU, Softmax	- Data augmentation - ES on MAE/L1 loss - Dropout - Class weighting	AdaDelta	-
Liu, L. [60]	U-Net (Semantic)	Dice	Self-gated soft attention / hybrid	ReLU, Sigmoid	- Data augmentation - Dropout	Adam	150
Moon, H. [50]	U-Net (Semantic)	BCE	None	ReLU, Sigmoid	-	Adam	200
Zhang, R. [19]	DenseNet (Semantic)	Dice	None	ReLU, Softmax	- Data augmentation - Weight decay - Learning rate adjust.	SGD	2000
Wong, K. [48]	U-Net (Semantic)	Dice	None	ReLU, ?	Data augmentation	-	-
Khezrpour, S. [79]	U-Net (Semantic)	Dice	None	ReLU, Sigmoid	- Data augmentation - ES on validation loss	Adam	-

 doi:10.20944/preprints202501.2233.v1

Hu, X. [90]	U-Net + ResNet (Semantic)	Focal	None	ReLU, Sigmoid	- Data augmentation - Class weighting	Adam	1500
Gheibi, Y. [37]	U-Net + ResNet (Semantic)	Custom	None	ReLU, Sigmoid	- Data augmentation - Weight decay - Dilution	Adam	-
Kumar, A. [110]	U-Net (Semantic)	BCE-Dice	None	ReLU, Softmax	- Data augmentation - Dropout - ES on validation set - Learning rate adjust.	Adam	200
Liu, L. [16]	U-Net + ResNet (Semantic)	Custom	None	Leaky ReLU, Sigmoid	Data augmentation	-	70
Zhao, B. [71]	CNN (Semantic)	BCE	Squeeze-excitation / channel	ReLU, Sigmoid	- Data augmentation - ES on validation loss	RAdam	-
Liu, C. [27]	U-Net (Semantic)	BCE-Dice	Dual attention gates / hybrid	SeLU (Self-normalized), Sigmoid	- Weight decay - ES on training & val. - Learning rate adjust. - Class weighting	Adam	200
Karthik, R. [44]	U-Net (Semantic)	Dice	Attention gates / spatial	ReLU, Sigmoid	Data augmentation	-	150
Liu, L. [95]	U-Net + DenseNet (Semantic)	CE-Dice	None	ReLU, Sigmoid	- Data augmentation - Dropout	Adam	8

 doi:10.20944/preprints202501.2233.v1

Aboudi, F. [59]	U-Net (Semantic)	CE	None	ReLU, Sigmoid	Data augmentation	Adam	100
Pinto, A. [97]	U-Net (Semantic)	Dice	None	-	-	Adam	-
Choi, Y. [96]	U-Net + CNN + ResNet (Semantic)	CE-Dice	None	ReLU, Softmax	- Data augmentation - Weight decay - Dropout *- ES on ?*	Adam	-
Kim, Y. [63]	U-Net (Semantic)	Dice	None	ReLU, Sigmoid	-	Adam	1000
Woo, I. [25]	U-Net + DenseNet (Semantic)	-	Squeeze-excitation / channel	ReLU, Sigmoid	-	-	-
Lee, A. [26]	U-Net (Semantic)	Dice	Squeeze-excitation / channel	ReLU, Sigmoid	-	-	-
Lee, S. [81]	U-Net (Semantic)	Dice	None	ReLU, Sigmoid	ES on validation loss	Adam	-
Karthik, R. [101]	U-Net (Semantic)	Dice-CE + Softmax-CE	Multi-residual attention / hybrid	ReLU, Softmax	- Data augmentation - Masked dropout - Dropout - Learning rate adjust.	Adam	150
Zhang, L. [99]	U-Net (Semantic + Instance)	CE	None	ReLU, Softmax	- Data augmentation - Momentum - Weight decay	SGD	-
Ou, Y. [103]	U-Net (Semantic)	BCE	None	ReLU, Softmax	-	RMSprop	100
Vupputuri, A. [102]	U-Net (Semantic)	BCE	Multi-path attention / hybrid (includes	Leaky ReLU, Softmax	- ES on validation set - Dropout	Adam	30

 doi:10.20944/preprints202501.2233.v1

			self-attention)				
Abdmoulah, N. [64]	U-Net (Semantic)	CE	None	ReLU, Sigmoid	Data augmentation	Adam	20
Duan, W. [98]	CNN + ResNet (Semantic)	Dice-CE	None	PReLU, Softmax	Data augmentation	Adam	600
Lucas, C. [100]	U-Net (Semantic)	Soft QDice	None	ReLU, Sigmoid	Data augmentation	Adam	100
Nazari-Farsani, S. [45]	U-Net (Semantic)	BCE-Volume-MAE-Dice	Attention gates / spatial	ReLU, Sigmoid	- Data augmentation - Class weighting - Dropout	Adam	80
Wei, Y. [49]	U-Net + ResNet (Semantic)	Focal Tversky	None	ReLU, Softmax	- Data augmentation - Class weighting - Learning rate adjust.	Adam	150
Li, C. [72]	U-Net (Instance)	CE	None	ReLU, Sigmoid	- Data augmentation - Class weighting	SGD	200
Liu, Z. [67]	CNN + ResNet (Semantic)	Dice	None	ReLU, Sigmoid	- Data augmentation - Weight decay - Learning rate adjust.	Adam	500
Cornelio, L. [58]	U-Net (Semantic)	Dice	None	ReLU, Sigmoid	- Dropout - Weight decay	Adam	50
Yu, Y. [46]	U-Net (Semantic)	BCE-Volume-MAE-Dice	Attention gates / spatial	ReLU, Sigmoid	- Data augmentation - Class weighting	Adam	120

 doi:10.20944/preprints202501.2233.v1

					- Dropout		
Wu, Z. [87]	U-Net + MLP (Semantic)	Dice + Boundary	Multi-head self-attention / hybrid (includes self-attention)	ReLU, Softmax	- Weight decay - Learning rate adjust.	AdamW	35
Guerrero, R. [34]	U-Net + ResNet (Semantic)	CCE	None	ReLU, Softmax	- Data augmentation - Class weighting - Learning rate adjust.	Adam	-

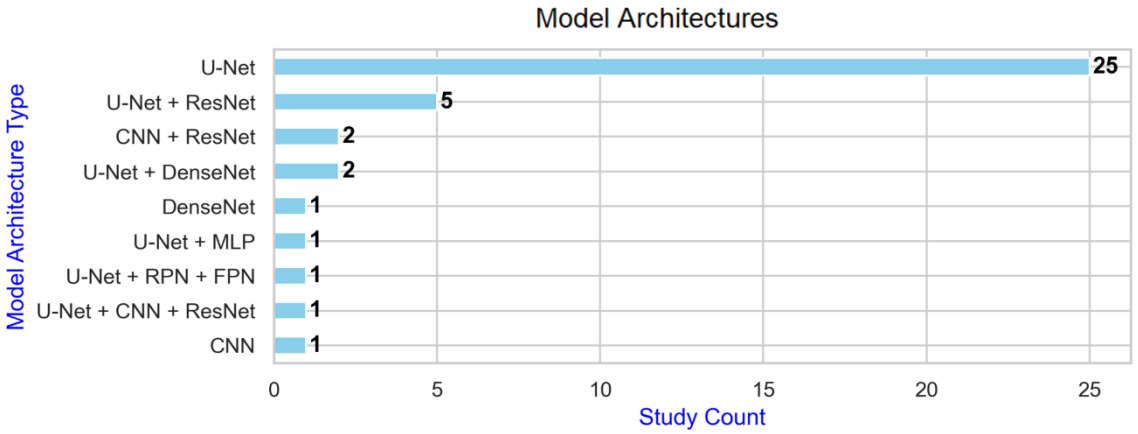

Figure 7. Model architecture types.

4.6. Performance and Generalisability

As Table 6 shows, the most used performance metrics across the studies reviewed were the overlap metrics Dice, Recall, and Precision, as well as the Hausdorff distance [61]. Six papers only used one single metric. To comparatively evaluate the models according to their performance, we assigned a generalisability score to each of the included studies based on sample representativeness – considering sample size, number of study sites, gender equality, age range, length of the data collection period, number of scanners, external validation performed –, ground-truth data, and access to clean code (Table 6, third column from right to left). Liu C. et al.'s algorithm [27] was deemed "highly" generalisable, whereas 19 algorithms had "low" generalisability. Plotting the reported performance against the generalisability scores obtained revealed that Dice and generalisability scores were positively correlated (Supplementary Figure S4a).

Only six papers analysed segmentation performance in relation to lesion size (i.e., on small versus large lesions), and in four of them, accuracy on small lesions was lower or significantly lower (Figure 8b). As shown in Supplementary Figure S4c, lesion volume ranges differed substantially between studies, and all cases with low mean Dice (< 0.6) (10 studies) reported low mean lesion volumes (< 40ml), while all cases with higher lesion volumes (> 60ml) (4 studies) reported high Dice

 doi:10.20944/preprints202501.2233.v1

scores (> 0.68). In other words, segmentation performance was generally better when lesions were larger.

As shown in Figure 8a, Dice scores were above the overall mean and relatively consistent across T2-WI, T1-WI, and FLAIR imaging modalities (mean Dice around 0.7), while PWI exhibited lower-than-average performance (mean Dice 0.38). Only for DWI did all the data points fall within the IQR (between 25th-75th percentiles), as outliers with below-average Dice scores were observed for FLAIR (three), T1WI (two) and T1WI (one). Additionally, the lower half of the IQR (25th-to-50th percentile) was substantially wider than the upper half (50th-to-75th percentile) for DWI, whereas the opposite pattern appeared in the IQR for PWI.

We also saw a positive correlation between spatial resolution and reported segmentation performance (Supplementary Figure S4b). Seven studies performed external validation of their models on unseen data. Most of them obtained higher Dice values on the study's test set than on the external validation set. We also observed a positive correlation between sample size and segmentation performance. Also, single-centre studies showed better performance (mean Dice 0.71) than multi-centre studies (mean Dice 0.6).

Dice scores were much higher for studies using ISLES-2015 (mean Dice > 0.7) or proprietary datasets (mean Dice 0.72), than when using ISLES-2017 (mean Dice 0.38) or DEFUSE (mean Dice 0.52). When attention-based networks were deeper, or when U-Nets were deeper, Dice scores were higher (Supplementary Figure S4d). The mean Dice was also higher when attention was used (0.71 versus 0.6 if not used).

Models using focal loss heavily under-performed, while those using learning rate adjustment over-performed. There was negative correlation between Dice scores and numbers of epochs used. Interestingly, only one of the algorithms that used a relatively high number of epochs was also using early stopping regularization, which means that for all the others, the full (high) amount of epochs was used during training, hence substantially increasing the probability of overfitting.

Table 6. Performance and generalisability data (see corresponding summary graphs in the **Supplementary Data B**).

First author	Dice	Precision	Recall	Hausdorff distance	Lesion size-based results	General-izability	Train time	Training library and infrastructure
	** Only scores reported on test sets are extracted* ** When scores are reported per input dataset, the average score is provided* ** Format: mean score ± standard deviation*				*Results as reported based on lesion size*			
Karthik, R. [42]	0.701	-	-	-	N	L	7h30	**CPU**: 3.6GHz QuadCore Intel Gen7 **RAM**: 32GB **GPU**: Nvidia Quadro P4000 **Library**: Keras/TensorFlow
Gómez, S. [23]	0.36 ± 0.21	0.42 ± 0.25	0.48 ± 0.29	-	N	L	-	-

 doi:10.20944/preprints202501.2233.v1

Olivier, A. [70]	0.703 ± 0.2	-	-	-	**Sensitivity**: S (<20mL): 0.987 L (>=20mL): 0.923 **Specificity**: S (<20mL): 0.923 L (>=20mL): 0.987	M	-	**GPU**: Nvidia Tesla K80 **Library**: Keras/TensorFlow
Clèrigues, A. [69]	0.715 ± 0.205	0.735 ± 0.25	0.745 ± 0.18	27.7 ± 21.45	N	M	-	**CPU**: Intel CoreTM i7-7800X **OS**: Ubuntu 18.04 **RAM**: 64GB **GPU**: Nvidia Titan X (12GB) **Library**: Torch
Liu, L. [60]	0.764	-	0.944	3.19	N	M	-	-
Moon, H. [50]	0.737 ± 0.32	0.758	0.755	22.047	Relation Dice-lesion size: Observed R2=0.195	L	24h	**OS**: Centos7 **GPU**: 4 x Nvidia Quadro RTX 8000 **Library**: Keras/TensorFlow
Zhang, R. [19]	0.791	0.927	0.782	-	N	M	6h23	**CPU**: Intel Core i7-4790 3.60GHz **RAM**: 16GB **GPU**: Nvidia Titan X **Library**: PyTorch
Wong, K. [48]	0.84 ± 0.03	0.84 ± 0.03	0.89 ± 0.03	-	N	M	-	-
Khezrpour, S. [79]	0.852	0.998	0.856	-	N	L	-	**GPU**: Google Cloud Compute (K80) **Library**: Keras/TensorFlow
Hu, X. [90]	0.30 ± 0.22	0.35 ± 0.27	0.43 ± 0.27	-	N	L	-	**GPU**: 4 x Nvidia Titan Xp
Gheibi, Y. [37]	0.792	-	-	-	N	M	1h27	**GPU**: Nvidia Tesla P100 **Library**: Keras

 doi:10.20944/preprints202501.2233.v1

Kumar, A. [110]	-	0.633 ± 0.213	0.653 ± 0.223	-	N	M	11h45	**CPU**: 2x Intel Xeon Silver 4114 (2.2GHz, 10C/20T) **RAM**: 192GB **GPU**: Nvidia Tesla V100 PCIe **Library**: Keras/TensorFlow
Liu, L. [16]	0.817	-	-	1.92	N	M	0h36	**Library**: Keras/TensorFlow
Zhao, B. [71]	0.699 ± 0.128	0.852	0.923	-	Dice: S: 0.718 (0.12) L: 0.689 (0.222)	M	-	**CPU**: Intel Core i7-6800K **RAM**: 64GB **GPU**: Nvidia GeForce 1080Ti **Library**: PyTorch
Liu, C. [27]	0.76 ± 0.16	0.83 ± 0.17	0.73 ± 0.19	-	Dice: S (<1.7ml): 0.68 (0.19) M (≥1.7&<14ml) : 0.75 (0.14) L (≥14ml): 0.83 (0.10)	H	-	**CPU**: Intel Core E5-2620v4 (2.1GHz) **GPU**: 2 x Nvidia Titan XP **Library**: Keras/TensorFlow
Karthik, R. [44]	0.7535	-	-	-	N	L	34h04	**CPU**: 3.6GHz QuadCore Intel (Gen 7) **RAM**: 32GB **GPU**: Nvidia Quadro P4000 **Library**: Keras/TensorFlow
Liu, L. [95]	0.68 ± 0.19	-	-	39.975 ± 27.95	N	M	-	-
Aboudi, F. [59]	0.558	0.998	-	-	N	L	-	**CPU**: Intel Core i5 8th gen **RAM**: 8GB **GPU**: Nvidia GeForce GTX 1050

 doi:10.20944/preprints202501.2233.v1

								Library: Keras/TensorFlow
Pinto, A. [97]	0.29 ± 0.21	0.23 ± 0.21	0.66 ± 0.29	41.58 ± 22.04	N	L	-	**GPU**: Nvidia GeForce GTX-1070 **Library**: Keras/Theano
Choi, Y. [96]	0.31	-	-	37.7	N	L	3h	**CPU**: 2 x Intel Xeon CPU E5-2630 v3 (2.4GHz) **GPU**: 4 x Nvidia GeForce GTX TITANX **Library**: Keras
Kim, Y. [63]	0.6 ± 0.23	-	-	-	Dice: > 0.75 for lesion volumes > 70mL	L	20h	**CPU**: Intel Xeon Processor E5-2680 (14 CPU, 2.4 GHz) **OS**: Ubuntu Linux 14.04 SP1 **RAM**: 64GB **GPU**: Nvidia GeForce GTX 1080 **Library**: TensorLayer
Woo, I. [25]	0.858 ± 0.0734	-	-	-	Dice: - S (<10mL): 0.82 - L (>10mL): 0.89	L	-	-
Lee, A. [26]	0.854 ± 0.008	0.845	0.995	-	N	L	-	-
Lee, S. [81]	0.422 ± 0.277	0.48 ± 0.308	0.467 ± 0.32	-	Dice: S (<10mL): 0.377 L (>10mL): 0.607	M	52h30	**CPU**: Xeon Processor E5-2650 v4 (Intel) **GPU**: Nvidia Titan X **Library**: Keras/TensorFlow
Karthik, R. [101]	0.775	0.751	0.801	-	N	L	-	**CPU**: 4 cores **OS**: Ubuntu 16.04 **RAM**: 32GB

 doi:10.20944/preprints202501.2233.v1

								GPU: 2 x Nvidia Tesla P100 **Library**: PyTorch
Zhang, L. [99]	0.433	-	0.356	-	N	L	-	**GPU**: Nvidia GeForce GTX 1080 Ti **Library**: Keras/TensorFlow
Ou, Y. [103]	0.865	0.894	0.818	-	N	M	4h	**GPU**: 4 x Nvidia Quadro RTX 6000 **Library**: PyTorch
Vupputuri, A. [102]	0.71	-	0.897	-	N	M	-	**GPU**: Nvidia Tesla K80
Abdmouleh, N. [64]	0.71 ± 0.11	-	-	-	N	L	-	-
Duan, W. [98]	0.677 ± 0.165	-	-	85.462 ± 14.496	N	M	-	**GPU**: Nvidia GTX 1080 Ti **Library**: PyTorch
Lucas, C. [100]	0.35	0.52	0.35	21.48	N	L	-	**GPU**: Nvidia Titan Xp (12GB) **Library**: PyTorch
Nazari-Farsani, S. [45]	0.5	-	0.6	-	N	M	-	-
Wei, Y. [49]	0.828	-	-	-	Dice: S (<769 pixels): 0.761 L (>769): 0.83	M	-	-
Li, C. [72]	-	-	-	38.27mm	N	L	-	-
Liu, Z. [67]	0.658	0.61	0.6	51.04	N	M	-	**CPU**: Intel Core i7-7700K **RAM**: 48GB **GPU**: Nvidia GeForce 1080Ti **Library**: Keras/TensorFlow
Cornelio, L. [58]	0.34	-	-	-	N	L	5h	**OS**: Ubuntu v.16.04.3 **GPU**: Nvidia GeForce GTX

 doi:10.20944/preprints202501.2233.v1

								Library: Keras/TensorFlow
Yu, Y. [46]	0.53	0.53	0.66	-	N	M	35h	**GPU**: Nvidia Quadro GV100 & Nvidia Tesla V100-PCIE **Library**: Keras/TensorFlow
Wu, Z. [87]	0.856	0.883	0.854	27.34	N	M	0h21	**GPU**: 6 x Nvidia Tesla 4s **Library**: PyTorch
Guerrero, R. [34]	0.4±0.252	-	-	-	N	L	-	**Library**: Lasagne/Theano

 doi:10.20944/preprints202501.2233.v1

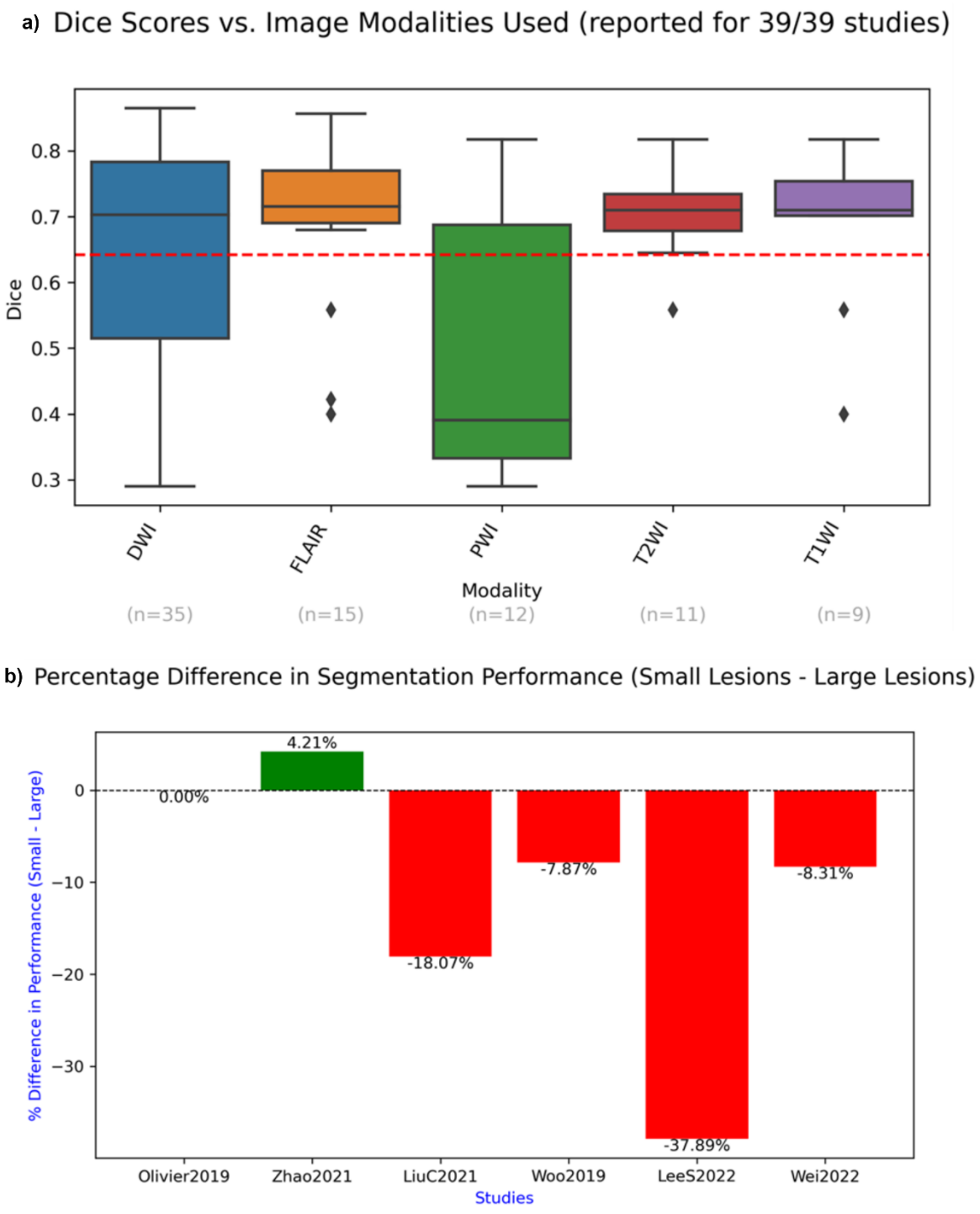

Figure 8. Impact of different MRI modalities on the accuracy of lesion segmentation. **a)** Box plot showing the correlation between Dice scores and imaging modalities used; **b)** Percentage difference in lesion segmentation performance for small versus large lesions, calculated as (small lesion performance - large lesion performance) relative to large lesion performance. Positive values indicate better performance on small lesions, while negative values indicate better performance on large lesions.

4.7. Reported Dice Scores and Segmentation Quality

We explored whether the reported Dice scores are a legitimate indicator of segmentation quality in this review. For this we generated a forest plot using the data from the 17 papers that reported their Dice along with their standard deviation (Figure 9). In this analysis, the percentage of variation across studies due to heterogeneity rather than chance (I^2) was 22.08%.

 doi:10.20944/preprints202501.2233.v1

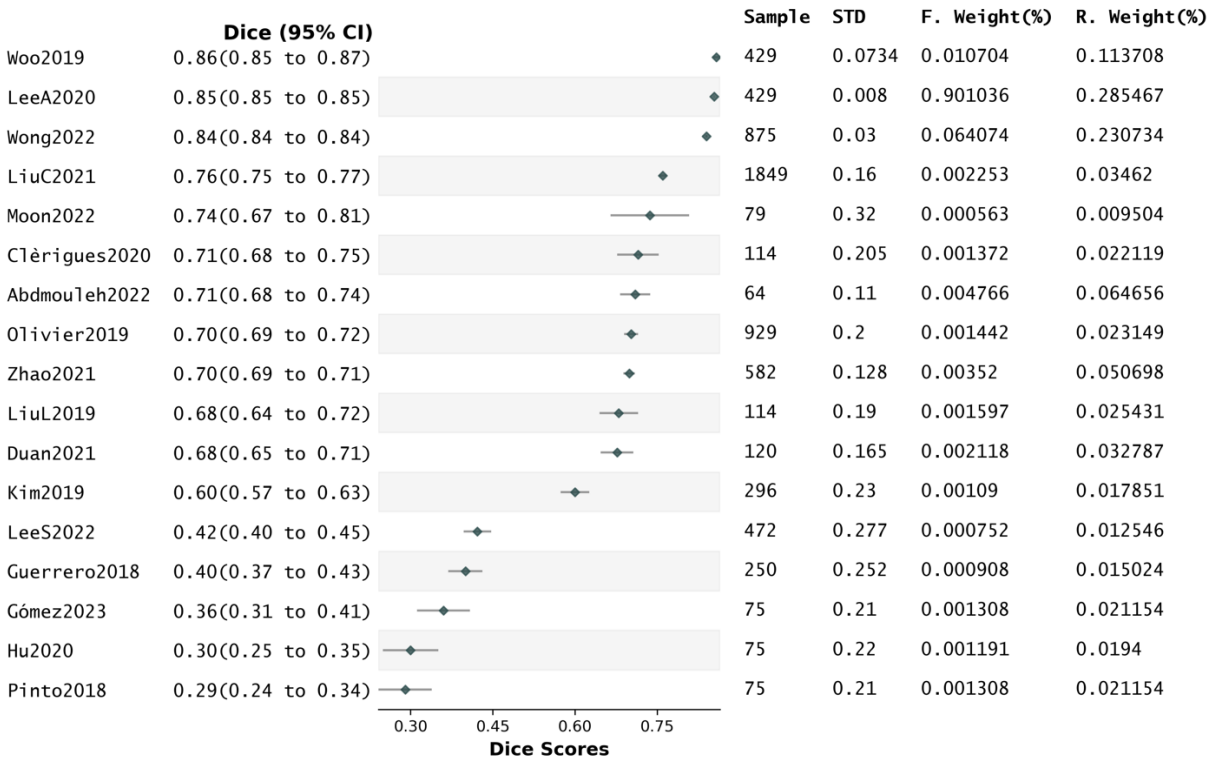

Figure 9. Forest plot related to the whole group analysis.

We also conducted a sensitivity analysis using Precision scores as effects size instead of Dice scores. This analysis involved only eight studies, which reported their precision scores along with standard deviations. But in this analysis, I^2 was 8.49%, indicating a reduced level of heterogeneity between studies, therefore precluding us to derive conclusions from it (Supplementary Figure S5).

Funnel plots and Egger's tests (Supplementary Figures S6 and S7) conducted using the Dice scores reported by the included studies indicated the presence of publication bias in favour of studies reporting high values of this metric.

4.8. Influence of Attention on Dice Scores

We conducted a subgroup analysis to evaluate the association between attention mechanisms and Dice scores. The resulting forest plot is shown in Figure 10.

 doi:10.20944/preprints202501.2233.v1

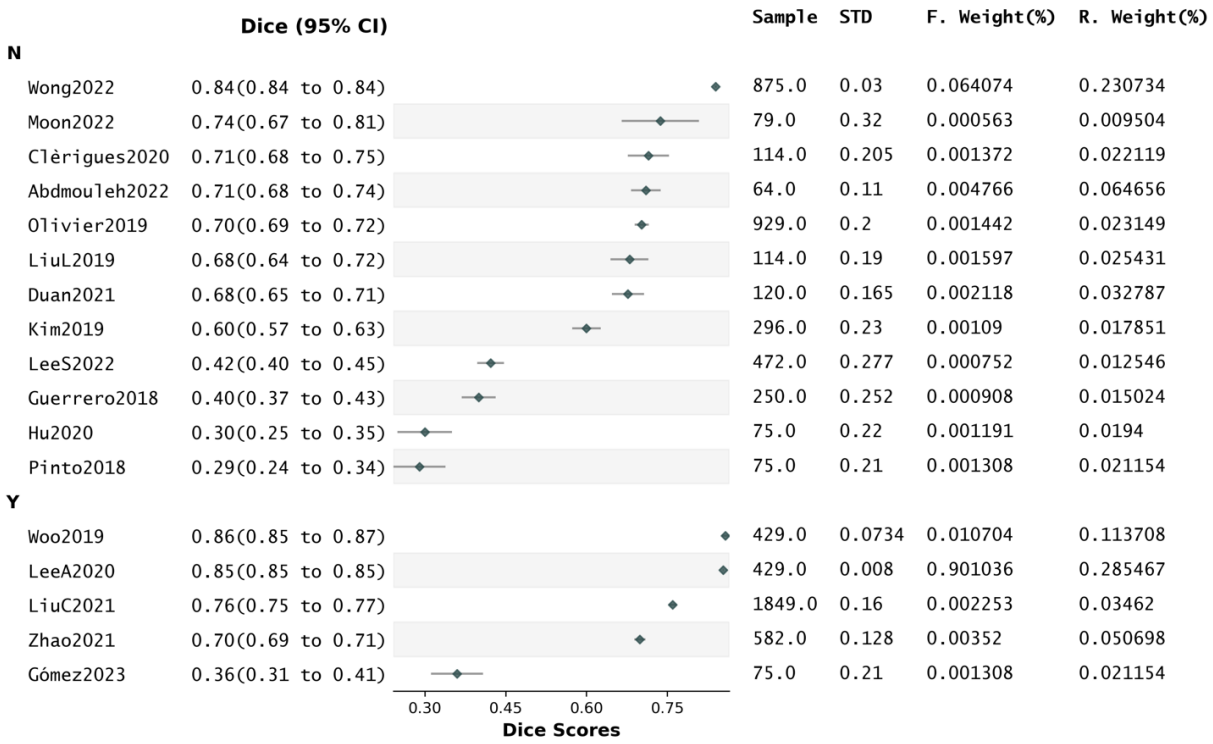

Figure 10. Forest plot related to the subgroup analysis.

There were no statistically significant differences in effect sizes between the groups. The subgroup "with attention" indicated moderate heterogeneity in I^2 (31.63%) and a very high Z-stat (39.03, $p<0.001$), suggesting a substantially large overall effect. While this implies that the presence of attention may enhance segmentation performance, the small number of studies in this subgroup (five) limits the conclusiveness of this result. In contrast, the subgroup "without attention" comprised 12 studies, showing significant heterogeneity in the Q-stat ($Q=19.93$, $p=0.05$) and in I^2 (39.80%). Despite the absence of attention, a significant overall effect was also observed ($z=5.27$, $p<0.001$). This suggests that when attention is not used, the Dice scores differ between studies.

Further meta-regression analysis to assess the statistical significance of the relationship between "attention mechanisms" and "Dice scores" (Supplementary Figure S8) revealed that 8.1% of the variance in Dice scores was explained by the presence of attention (R-squared: 0.081), but the slope indicating the change in Dice associated with the presence of attention was not statistically significant (0.117, $p=0.27$, 95% CI of the slope [-0.100,0.334]). This indicates that from the literature analysis we cannot conclude that the presence of attention has a significant impact on the likelihood of high Dice.

4.9. Risk of Bias Assessment

After assessing the possibility of biases in the included studies, 31 studies scored "GOOD", and eight scored "FAIR" in the NIH study QA (Supplementary Data C). Although these results are positive, we identified cases of potential *spectrum bias* [107], mostly due to the following factors: acute stroke studies were more represented than subacute (30 versus 16), exposure was often only assessed once (i.e., no follow-up scans) (24 studies), variance and effect estimates were not both provided (21 studies), few experiments were conducted to assess the different levels of exposure related to the outcome (11 studies), period of data collection was relatively short (10 studies), study population was poorly defined (3 studies), and the age range of participants was not always consistent (e.g., Kim et al. [63] only included patients between 58-79 years old).

We also noticed cases of *selection bias*. Multiple studies used the same ISLES datasets to evaluate the performance of their segmentation methods. This, although advantageous (e.g., cost effective, allows comparability), introduces selection bias. These were also studies where males were over-represented in the sample.

 doi:10.20944/preprints202501.2233.v1

Also, ground-truth data were most often obtained by manually refining semi-automatic segmentations (e.g., thresholding followed by region-growing), which introduces *observer bias*. Sixteen studies did not provide information about labelling criteria, so it is unclear whether observer bias was present in those.

We identified two other forms of bias: *verification bias* in 10 studies, where only one expert did the labelling of ground-truth images, and *measurement bias*, as mean Dice scores on ISLES-2017 were generally much lower than those on ISLES-2015, and when segmentation performance was reported for small versus large lesions, the definition of a small and a large lesion (in ml) was not consistent across studies.

4.10. Pilot Analysis

The best performing model was "UResNet50" on DWI (single-modality approach), using a weighted compound loss (BCE = 0.3 + Dice = 0.7), with a Dice score on the validation set of 0.692 ± 0.132 (Table 7).

Table 7. Results from the pilot analysis.

	Mean Dice Score (± STD)			
	DWI	DWI	DWI + FLAIR + T1WI + T2WI	DWI + FLAIR + T1WI + T2WI
UResNet50	Train	Validation	Train	Validation
BCE=0.3 + Dice=0.7	0.911 ± 0.11	0.692 ± 0.132	0.908 ± 0.041	0.675 ± 0.128
BCE=0.5 + Dice=0.5	0.893 ± 0.102	0.610 ± 0.055	0.884 ± 0.318	0.619 ± 0.301
BCE=0 + Dice=1	0.902 ± 0.205	0.625 ± 0.306	0.886 ± 0.16	0.608 ± 0.04
UNet	Train	Validation	Train	Validation
BCE=0.3 + Dice=0.7	0.843 ± 0.322	0.556 ± 0.083	0.838 ± 0.072	0.570 ± 0.159
BCE=0.5 + Dice=0.5	0.829 ± 0.031	0.521 ± 0.29	0.836 ± 0.2	0.547 ± 0.234
BCE=0 + Dice=1	0.837 ± 0.202	0.560 ± 0.105	0.842 ± 0.085	0.555 ± 0.18
AG-UResNet50	Train	Validation	Train	Validation
BCE=0.3 + Dice=0.7	0.907 ± 0.121	0.676 ± 0.222	0.909 ± 0.177	0.664 ± 0.313
BCE=0.5 + Dice=0.5	0.899 ± 0.06	0.642 ± 0.176	0.873 ± 0.096	0.630 ± 0.269
BCE=0 + Dice=1	0.893 ± 0.19	0.669 ± 0.091	0.877 ± 0.231	0.631 ± 0.164
AG-UNet	Train	Validation	Train	Validation
BCE=0.3 + Dice=0.7	0.829 ± 0.258	0.522 ± 0.142	0.817 ± 0.109	0.536 ± 0.22
BCE=0.5 + Dice=0.5	0.793 ± 0.2	0.518 ± 0.207	0.802 ± 0.163	0.515 ± 0.082
BCE=0 + Dice=1	0.797 ± 0.32	0.529 ± 0.099	0.784 ± 0.27	0.498 ± 0.105

The second best was "AG-UResNet50" (0.676 ± 0.222), with a single-modality approach, and using the same compound loss (BCE = 0.3 + Dice = 0.7).

Experiments with "UNet" and "AG-UNet" generated relatively poor Dice scores. Performance was better in single-modality experiments. Abdmouleh et al. [64] made the same test on the same dataset, but they achieved quasi-equal performance in their DWI-only and multi-modal experiments (Dice 0.71). Performance was also better when using compound loss "BCE = 0.3 + Dice = 0.7" versus the other two types. The 12 experiments using attention and the 12 not using attention yielded similar average Dice scores.

Average training times for UResNet50 was 5h 43min, for U-Net it was 5h 31min, for AG-UResNet50 it was 6h 15min, and for AG-UNet it was 5h 55min. Multi-modal experiments took longer

 doi:10.20944/preprints202501.2233.v1

to train in all cases (~3 hours longer each time). Same was true for attention-based experiments (~30 minutes longer each time).

5. Discussion

5.1. Systematic Review and Meta-analysis

Although our review protocol did not have age restriction, samples never included patients below 18 years old. This stresses the lack of research in paediatric stroke, which may be due to multiple factors, e.g., delayed identification of stroke, numerous stroke aetiologies and risk factors in children, and limited imaging data [65].

The underrepresentation of females in studies, on the other hand, can be partially explained by the difficulty of diagnosing females with stroke, due to factors such as higher proportion of stroke mimics (e.g., migraine), pre-stroke disability, or neglect of symptoms among females [66]. These uneven distributions of gender and age data can affect the universality of our research outcomes.

Most studies focused exclusively on minor-to-moderate stroke cases with focus on acute stroke, since DWI and FLAIR are able to show high signal in AIS-affected brain areas, whereas signal begins to diminish gradually towards the subacute stage, often leading to lower MRI sensitivity for stroke identification [53]. Such differences in MRI signal between subacute and acute lesions give the idea that combining acute and subacute cases in one single dataset, as seen in Liu C. et al. [27] and Liu Z. et al. [67], might require highly trained observers to manually delineate the lesions (i.e., generate the reference labels).

We also noticed relatively small sample sizes across studies, which is not new in AIS research [68]. Data augmentation is a common way to mitigate this issue, and Clèrigues et al. [69] proposed a novel "symmetric modality augmentation" technique, which leveraged learned features based on the symmetry of brain hemispheres. Other ways to deal with small sample sizes include active learning (e.g., Olivier et al. [70]), semi-supervised learning using weakly labelled data (e.g., Zhao et al. [71]), or transfer learning (e.g., Li et al. [72] used TernausNet [73] which was pre-trained on ImageNet [74]).

Several studies used high spatial resolution images to capture more fine-grained features from the data and improve segmentation performance on small lesions. Other deepened their networks further to collect more nuanced features, but the higher the number of down-sampling operations, the lower the resolution of the feature maps, to a point where reconstructing lesions in the up-sampling path becomes virtually impossible. Furthermore, risks of overfitting/over-learning increase substantially when networks are deeper, especially in absence of skip connections.

More generally, using 3T magnetic field strength, as done by 34/39 studies, can also help with small lesions, as it offers better signal-to-noise ratio and spatial resolution versus 1.5T, and it reduces imaging artifacts by offering more uniform B1 inhomogeneity [75].

Most studies used DWI, known as the gold standard for early stroke detection [76], and many used T1-WI, a staple in subacute stroke research [77], T2-WI, PWI, or FLAIR. PWI was frequently applied to detect the ischaemic penumbra [57], and most used FLAIR as it offers enhanced lesion clarity by suppressing CSF details [78]. For instance, Khezrpour et al.'s U-Net used only FLAIR and got very high accuracy [79]. ADC maps were also often used with DWI for more robust ground-truth data, as lesions appear simultaneously hyperintense on DWI and hypointense on ADC in early stroke stages.

The impact of using different imaging modalities (i.e., T1-WI, T2-WI, DWI, PWI, FLAIR) on lesion segmentation accuracy was also observed, as each modality may highlight distinct pathological features, which may, in turn, influence algorithm performance.

DWI-PWI mismatch [80] was commonly used to create ground-truth sets (e.g., Lee S. et al. [81]), since PWI identifies penumbral tissue, while DWI delineates the core infarct (i.e., areas of restricted water diffusion [77]). Despite its utility though, DWI-PWI mismatch analysis remains challenging. Establishing clear imaging boundaries for recoverable tissue is not straight-forward [77]. Large perfusion abnormalities may be observed in patients without corresponding clinical deficits [82].

 doi:10.20944/preprints202501.2233.v1

There is no universally defined mismatch ratio, although Kakuda et al. tried to define one [83] DWI-FLAIR mismatch, on the other hand, is mostly used for TSS assessment in hyper-acute-to-early-acute stage [84]. Combining both mismatch analyses can definitely help experts effectively delineate stroke lesions.

Many argue that using 3D images is crucial for DL-based stroke lesion segmentation, but (i) few methods address the associated computational challenges [85], which explains why the majority of retained studies used 2D images, and (ii) 3D models proved to be good at segmenting large organs, but are less "established" in stroke lesion segmentation [10].

Cutting 3D images into 3D patches (i.e., patch-wise training) is a way to mitigate both the computational challenges, by reducing memory overhead [13], and the small lesions challenge, by forcing the model to focus on a smaller area of the entire image. That explains why eight out of ten 3D studies in this review have used patch-wise training.

On the other hand, the majority of studies that used ISLES-2015/2017 have processed those as 2D images, mainly due to their low-resolution when processed as 3D (slice thickness: 5mm). However, it was surprising to see so many 3D models use low resolution images, since the whole point of 3D models is to capture detailed information from images [86]. For instance, Zhang R. et al. [19] proposed a 3D model that captured both low-level local features and high-level ones, but they used low-resolution images.

Dice was the most used performance metric across studies, as (i) it is simple to interpret, (ii) it handles class imbalance, and (iii) its widespread use facilitates comparison between different methods. However, it remains an overlap metric that is prone to instability, especially with small lesions [87], and for an evaluation to be holistic, it must be accompanied by other types of metrics (e.g., surface-based, boundary-based, volume-based).

Dice scores were higher for single-center studies, but since too few of these studies performed external validation, we cannot exclude "over-adaptation" to the image acquisition protocol(s) from that one center, and therefore poor model generalisability.

Moving on to loss functions, CE loss quantifies the difference between two probability distributions (e.g., predictions and ground-truth), but it cannot handle class imbalance since each pixel/voxel contributes equally to the loss, and therefore the learning process may easily fall into a local optimal solution [88].

Focal loss is an adaptation of CE loss that introduces a modulating factor aimed at down-weighting the impact of well-classified examples [89], but since "lesion" is already the minority class in our case, focal loss overly penalizes correctly classified lesion pixels, which explains the very bad performance of studies using it (e.g., Hu et al.'s Brain SegNet [90]).

Generally, overlap-based loss functions (e.g., Dice loss) are more robust to data imbalance issues [89]. By penalizing false positives and false negatives differently, Dice loss indirectly encourages better performance on minority classes. However, despite its common usage, Dice loss has some limitations [89]: it fails to capture the distance between non-overlapping but close lesions, overlooks precise contour details (combining it with a boundary-based loss may help), and it disproportionately penalizes small lesions, especially in presence of large lesions, as opposed to distribution-based loss functions (e.g., CE loss) which have no such bias. A few custom loss functions have also been proposed to address class imbalance (e.g., Rachmadi et al.'s "ICI loss" [91], loss with data fusion [92]).

Since most studies were U-Net-based, they primarily performed semantic lesion segmentation. Perhaps the fact that only two studies did instance segmentation is linked to the difficulty of delineating individual lesions in presence of motion artefacts and irregular shapes [93,94], as shown by Wu et al. [87].

Meanwhile, several studies proposed quite innovative methods:

- Liu Z. et al. [67] proposed a ResNet and a global convolution network-based (GCN) encoder-decoder. Each modality was concatenated to a three-channel image, then passed as input image to a series of residual blocks. The output of each block was then passed to its corresponding up-sampling layer using a skip connection incorporating a GCN and a boundary refinement layer

 doi:10.20944/preprints202501.2233.v1

- Liu L. et al.'s "MK-DCNN" [95] consisted of two sub-DenseNets with different convolution kernels, aiming to extract more image features than with a single kernel by combining low and high resolution
- Three studies proposed "ensemble mechanisms" (i.e., different networks that process the same input in parallel) in order to reduce overfitting, since sub-networks can learn different features from the data [13] and/or to decrease prediction variance (e.g., Choi et al. [96])
- Wu et al.'s W-Net [87] tackled variability in lesion shape by trying to capture both local and global features in input scans. A U-Net first captures local features, which then go through a Boundary Deformation Module, then finally through a Boundary Constraint Module that uses dilated convolution to ensure pixels neglected in previous layers can also contribute to the final segmentation
- Pinto at al. [97], Duan et al. [98] and Zhang L. et al. [99] proposed "information fusion mechanisms" that effectively fuse different features either from multiple modalities, or multiple plane views, thus improving their models' ability to capture intricate lesion features
- Lucas et al. [100] added to their U-Net skip connections around each convolution block, besides those linking encoder-decoder layers

The main purpose of attention mechanisms is to address the loss of information during down-sampling and up-sampling operations. Self-attention was often used across studies, since it allows the model to capture global dependencies within the input data, which can help in identifying subtle features that span across larger regions.

Overall, there were several interesting implementations, or pseudo-implementations, of attention:

- Karthik et al. [101] embedded multi-residual attention blocks in their U-Net, hence allowing the network to use auxiliary contextual features to strengthen gradient flow between blocks and prevent vanishing gradient issues
- Vupputuri et al. [102] used self-attention through multi-path convolution, aiming to compensate for information loss, while using weighted average across filters to provide more optimal attention-enabled feature maps
- Ou. et al. [103] used lambda layers, which work by transforming intra-slice and inter-slice context around a pixel into linear functions (or "lambdas"), which are then applied to the pixel to produce enhanced features. As opposed to attention, lambdas do not give "weights" to pixels

I believe that it is only a coincidence that ResNet-based models never incorporated attention across reviewed studies, as numerous relevant publications combine ResNet with attention [104–106].

In terms of optimization methods, RMSProp can be effective in DL (e.g., Ou et al. [103]), as it is able to discard history from the extreme past and thus enable rapid convergence during training. However, Adam remains the most popular method as it incorporates momentum, which speeds up the optimization of model parameters, while performing bias corrections to improve the accuracy of gradient estimates during training. Also, Adam's default hyperparameters often work well in DL, mainly thanks to the adaptive learning rates which allow smooth parameter updates even in presence of noisy gradients.

While never performed, uncertainty quantification to obtain true network uncertainty estimates [107] is of utmost importance to promote the use of such algorithms in clinical practice, as it would allow physicians to assess when the network is giving unreliable predictions [6].

The generalisability of our studies was generally low, for issues that have already been highlighted above (e.g., small sample sizes, loose verification of labelled data), but researchers can

 doi:10.20944/preprints202501.2233.v1

easily improve the generalisability of their models by performing external validation, publishing their code, combining image acquisition protocols, and/or combining data from multiple centers.

Our risk of bias assessment yielded fairly good results. However, several instances of potential or actual bias warrant attention:

- Findings drawn from reported performance metrics (e.g., Dice) must be carefully interpreted, as performance depends on the quality of the data being used, which was variable across studies
- Results of this review may be skewed towards acute stroke (rather than subacute)
- Over-reliance on specific public datasets, which may have selection biases, may limit the generalisability of the research findings, as reported results may not fully represent all possible clinical scenarios
- Findings in terms of segmentation of small versus large lesions are slightly flawed, due to the various ways in which these two categories were defined across studies
- Data augmentation helped reduce overfitting by increasing the size of the training data, but effects of bias cannot be balanced-out by increasing the sample size by repetition [62]

Our whole group analysis included 17 studies, which is enough to consider findings meaningful [108]. The random-effects model worked better for us, which is aligned with the literature, where RE is considered a more natural choice than FE in medical research [109].

The most interesting finding resulted from the subgroup analysis. It is the uncertainty in the evidence that incorporating attention into DL architecture for AIS lesion segmentation improves model performance.

Meanwhile, the significant heterogeneity observed through these analyses may be linked to several factors, such as differences in image acquisition protocols (e.g., spatial resolution, scanners), patient populations (e.g., stroke stage, severity, aetiology), network architecture (e.g., U-Net, ResNet), model hyper-parameters, and more.

Therefore, when looking into ways to improve DL-based stroke lesion segmentation algorithms, our analysis suggests that **one might want to look at factors other than attention** (e.g., image quality, model architecture and complexity).

5.2. Pilot Analysis

The relatively high Dice scores obtained on training sets versus validation sets are likely caused by overfitting, partly due to the small sample size, despite efforts to mitigate this with data augmentation and pixel dropout.

We used the ISLES-2015-SISS dataset for this analysis. It is worthwhile noting that it may not sufficiently capture the variability across different populations and lesion types, and the limited sample diversity could limit the generalizability of the model across different demographics or lesion types. However, from the 39 publications analyzed, only 12 used this sample in the development of their proposed algorithm sometimes as part of a wider sample (5/12 publications) (Table 4). In terms of number of 3D volumes the sample is small, but we use a 2D model for which the number of input samples with image information multiplies the available data sources by a factor of approximately 100 considering only one dimension (e.g., considering horizontal-only or sagittal-only or coronal-only slices), but if slices in the three main imaging axes are considered, then the increase is three times that.

Not using attention yielded slightly better than using it. In this case, with a small sample size and a relatively deep network, **increasing the number of learnable parameters using attention gates might have accentuated the overfitting problem**. Complementing our analysis with additional performance metrics (e.g., HD, Accuracy, Precision) could further support this observation.

The fact that the single-modality approach (DWI-based) performed better than the multi-modal approach is counter-intuitive, since combining sequences has often led to an improved segmentation performance, as shown by Liu L. et al. [16] and Liu Z. et al. [67], who did the same comparison of

 doi:10.20944/preprints202501.2233.v1

approaches. However, it could be that specifically in the ISLES-2015-SISS dataset, the mix of image acquisition protocols across centers has introduced noise in the data, which was not properly removed during data pre-processing [57].

Compound loss (Dice + CE) outperformed Dice loss, as it was the case with Kumar et al.'s "CSNet" [110]. Since Dice loss is not suitable for small diffuse lesions, combining distribution-based loss with region-based loss has certainly helped.

UResNet50 addresses the challenge of distinguishing stroke lesions from other pathologies, which can vary by stage. Its effectiveness confers it potential to improve diagnostic accuracy and treatment planning for stroke patients, ultimately contributing to better clinical outcomes.

6. Study Limitations

- Only articles published in (or translated to) English that were accessible via institutional login were reviewed. Accordingly, relevant papers may have been missed
- Relevant papers may have been missed as a result of incongruences between search terms and article keywords in the various databases
- Since most of the included studies were not longitudinal, this review lacks an assessment of long-term patient outcomes, which is an essential factor in validating the clinical relevance and predictive value of segmentation algorithms
- While the review outlines the impact of lesion size on segmentation performance, the pilot analysis does not specifically assess how algorithms can be optimized for lesions of varying sizes

7. Conclusions and Future Works

While we included a fair number of studies in this review, the identified generalisability issues hinder the robustness of our findings. However, we were able to (i) identify the often subtle elements and configurations that can make a DL model perform better its AIS lesion segmentation task, and to (ii) demonstrate with confidence that attention mechanisms do not necessarily improve current DL architecture is AIS semantic lesion segmentation, and that other details such as model design were much more important.

We have compared multiple model artifacts (e.g., loss functions, optimization methods), discussing their potential impacts on segmentation performance. A more formal decision tree could complement our research, helping to (i) facilitate decision-making during model development, and (ii) enhance model transparency and trustworthiness in clinical settings.

In this review, algorithms were assessed solely based on performance (using Dice coefficients). A more comprehensive evaluation of their practical value could be conducted in future work by considering factors such as processing time and resource consumption.

More generally, further well-conducted and well-reported research is needed in this field, focusing on: (i) larger datasets, potentially by leveraging consortia such as the Human Connectome Project (https://www.humanconnectome.org/) or ENIGMA (https://enigma.ini.usc.edu/), (ii) higher-quality data, such as generating structured labels from radiologist reports [111], and (iii) longitudinal data to better assess how segmentation results impact patient treatment and prognosis.

Interpretability of algorithms must also improve, as today, computer scientists focus primarily on reaching higher levels of accuracy, while clinical researchers focus on verifying associations with patient outcomes [112]. For instance, deconvolution networks and guided back-propagation can explain the inner workings of DL networks [113,114].

Also, model fine-tuning remains time-consuming. Perhaps "Neural Architecture Search" will soon be a robust solution for automatic selection and parameterization of DL models [115].

At last, following the big leap DL took with the advent of GPU, many scientists are getting prepared for the next big leap, with quantum computing. Although this review did not focus on such

 doi:10.20944/preprints202501.2233.v1

technological advancements, the application of quantum algorithmic principles (e.g., running quantum operations on qubits) to ML has already begun [116], and expertise is being built for when quantum hardware will be commercially available. This may increase computing speed significantly.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org

Author Contributions: MB and MCVH conceived the topic, data extraction, and analyses of this systematic review, and planned the experiments linked to the pilot analysis, MB carried them out. MB and MCVH provided critical feedback and analysis, and contributed to the manuscript.

Funding: This work was funded by the University of Edinburgh (MB, MCVH), the Row Fogo Charitable Trust (Grant no. BRO-D.FID3668413) (MCVH), Dementias Platform UK 2, which receives funds from the UK Medical Research Council (MR/T033371/1), and the UK Dementia Research Institute at the University of Edinburgh (award number UK DRI-4002) through UK DRI Ltd, principally funded by the UK Medical Research Council, and additional funding partner the British Heart Foundation (MCVH, vascular group).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. MCVH is Specialty Chief Editor in Frontiers in Medical Technology.

List of Abbreviations

ADC, Apparent Diffusion Coefficient; **AIS**, Acute Ischemic Stroke; **AG**, Attention Gate; **BCE**, Binary Cross-Entropy; **BN**, Batch Normalization; **BOLD**, Blood Oxygenation Level Dependent; **CNN**, Convolution Neural Network; **CSF**, Cerebrospinal Fluid; **DenseNet**, Dense Convolutional Network; **DL**, Deep Learning; **DWI**, Diffusion-Weighted Imaging; **EHR**, Electronic Health Record; **ES**, Early Stopping; **FCN**, Fully-Convolutional Network; **FE**, Fixed-Effects; **FLAIR**, Fluid-Attenuated Inversion Recovery; **FPR**, False Positive Rate; **FNR**, False Negative Rate; **GCN**, Global Convolution Network; **HD**, Hausdorff's Distance; **HPC**, High Performance Computing; **IQR**, Interquartile Range; **MA**, Meta-Analysis; **ML**, Machine Learning; **MLP**, Multi-layer Perceptron; **MRI**, Magnetic Resonance Imaging; **NIH**, National Institute of Health; **NLP**, Natural Language Processing; **PRISMA**, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; **PWI**, Perfusion-Weighted Imaging; **QA**, Quality Assessment; **RE**, Random-Effects; **ReLU**, Rectified Linear Unit; **ResNet**, Residual Network; **SE**, Standard Error; **STD**, Standard Deviation; **T1-WI**, T1-Weighted Imaging; **T2-WI**, T2-Weighted Imaging; **TSS**, Time-Since-Stroke; **UoE**, University of Edinburgh; **WMH**, White Matter Hyperintensities.

References

1. Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Beaton, A. Z., Boehme, A. K., Buxton, A. E., Commodore-Mensah, Y., Elkind, M. S. V., Evenson, K. R., Eze-Nliam, C., Fugar, S., Generoso, G., Heard, D. G., Hiremath, S., Ho, J. E., … Martin, S. S. (2023). Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. Circulation, 147(8).
2. Saka, O., McGuire, A., & Wolfe, C. (2008). Cost of stroke in the United Kingdom. Age and Ageing, 38(1), 27–32.
3. Zhou, Y., Huang, W., Dong, P., Xia, Y., & Wang, S. (2021). D-UNet: A Dimension-Fusion U Shape Network for Chronic Stroke Lesion Segmentation. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(3), 940–950.
4. Hernandez Petzsche, M. R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.-L., Kofler, F., Ezhov, I., Robben, D., Hutton, A., Friedrich, T., Zarth, T., Bürkle, J., Baran, T. A., Menze, B., Broocks, G., Meyer, L., … Kirschke, J. S. (2022). ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific Data, 9(1), 762.

 doi:10.20944/preprints202501.2233.v1

5. Lo, E. H. (2008). A new penumbra: transitioning from injury into repair after stroke. Nature Medicine, 14(5), 497–500.
6. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.
7. Caceres, P. (2020). Introduction to Neural Network Models of Cognition (NNMOC). https://com-cog-book.github.io/com-cog-book/features/cov-net.html.
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org.
9. Abang Isa, A. M. A. A., Kipli, K., Mahmood, M. H., Jobli, A. T., Sahari, S. K., Muhammad, M. S., Chong, S. K., & AL-Kharabsheh, B. N. I. (2020). A Review of MRI Acute Ischemic Stroke Lesion Segmentation. International Journal of Integrated Engineering, 12(6).
10. Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2019). Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. Journal of Digital Imaging, 32(4), 582–596.
11. Ronneberger O., Fischer, P., Brox T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In J. and W. W. M. and F. A. F. Navab Nassir and Hornegger (Ed.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Pp. 234–241, Springer International Publishing.
12. Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift Für Medizinische Physik, 29(2), 102–127. https://arxiv.org/abs/1811.10052.
13. Liu, L., Cheng, J., Quan, Q., Wu, F.-X., Wang, Y.-P., & Wang, J. (2020). A survey on U-shaped networks in medical image segmentations. Neurocomputing, 409, 244–258.
14. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
15. Surekha, Y., Koteswara Rao, K., Lalitha Kumari, G., Ramesh Babu, N., & Saroja, Y. (2022). Empirical Investigations To Object Detection In Video Using ResNet-AN Implementation Method. Journal of Theoretical and Applied Information Technology, 100(10).
16. Liu, L., Chen, S., Zhang, F., Wu, F.-X., Pan, Y., & Wang, J. (2020). Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI. Neural Computing and Applications, 32(11), 6545–6558.
17. Veit, A., Wilber, M., & Belongie, S. (2016). Residual Networks Behave Like Ensembles of Relatively Shallow Networks.
18. Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV), 565–571.
19. Zhang, R., Zhao, L., Lou, W., Abrigo, J. M., Mok, V. C. T., Chu, W. C. W., Wang, D., & Shi, L. (2018). Automatic Segmentation of Acute Ischemic Stroke From DWI Using 3-D Fully Convolutional DenseNets. IEEE Transactions on Medical Imaging, 37(9), 2149–2160.
20. Diganta, M. (2020). Attention Mechanisms in Computer Vision: CBAM. Https://Blog.Paperspace.Com/Attention-Mechanisms-in-Computer-Vision-Cbam.
21. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 53, 197–207.
22. Takyar, A. (2021). How Attention Mechanism's Selective Focus Fuels Breakthroughs in AI. https://Www.Leewayhertz.Com/Attention-Mechanism/.
23. Gómez S, Mantilla D, Rangel E, Ortiz A, Vera DD, Fabio Martínez Carrillo. A deep supervised cross-attention strategy for ischemic stroke segmentation in MRI studies. Biomedical physics & engineering express. 2023 Apr 5;9(3):035026–6.
24. Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. https://arxiv.org/abs/1709.01507.

25. Woo, I., Lee, A., Jung, S. C., Lee, H., Kim, N., Cho, S. J., Kim, D., Lee, J., Sunwoo, L., & Kang, D.-W. (2019). Fully Automatic Segmentation of Acute Ischemic Lesions on Diffusion-Weighted Imaging Using Convolutional Neural Networks: Comparison with Conventional Algorithms. Korean Journal of Radiology, 20(8), 1275.
26. Lee, A., Woo, I., Kang, D.-W., Jung, S. C., Lee, H., & Kim, N. (2020). Fully automated segmentation on brain ischemic and white matter hyperintensities lesions using semantic segmentation networks with squeeze-and-excitation blocks in MRI. Informatics in Medicine Unlocked, 21, 100440.
27. Liu, C.-F., Hsu, J., Xu, X., Ramachandran, S., Wang, V., Miller, M. I., Hillis, A. E., Faria, A. v, Wintermark, M., Warach, S. J., Albers, G. W., Davis, S. M., Grotta, J. C., Hacke, W., Kang, D.-W., Kidwell, C., Koroshetz, W. J., Lees, K. R., Lev, M. H., … investigators, T. S. and V. I. (2021). Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke. Communications Medicine, 1(1), 61.
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. (2021, October). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;372(71).
29. Linares-Espinós E, Hernández V, Domínguez-Escrig JL, Fernández-Pello S, Hevia V, Mayor J, et al. (2018). Methodology of a systematic review. Actas Urol Esp (Engl Ed). 42(8):499-506.
30. Maier, O., Menze, B. H., von der Gablentz, J., Häni, L., Heinrich, M. P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H.-L., Havaei, M., … Reyes, M. (2017). ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis, 35, 250–269.
31. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2018). DRINet for Medical Image Segmentation. IEEE Transactions on Medical Imaging, 37(11), 2453–2462.
32. Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2018). Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. https://arxiv.org/abs/1802.06955.
33. Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P.-A. (2018). VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage, 170, 446–455.
34. Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M. C., Dickie, D. A., Wardlaw, J., & Rueckert, D. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. NeuroImage: Clinical, 17, 918–934.
35. Drozdzal, M., Chartrand, G., Vorontsov, E., Shakeri, M., di Jorio, L., Tang, A., Romero, A., Bengio, Y., Pal, C., & Kadoury, S. (2018). Learning normalized inputs for iterative estimation in medical image segmentation. Medical Image Analysis, 44, 1–13.
36. Jin, Q., Meng, Z., Sun, C., Cui, H., & Su, R. (2020). RA-UNet: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT Scans. Frontiers in Bioengineering and Biotechnology, 8.
37. Gheibi, Y., Shirini, K., Razavi, S. N., Farhoudi, M., & Samad-Soltani, T. (2023). CNN-Res: deep learning framework for segmentation of acute ischemic stroke lesions on multimodal MRI images. BMC Medical Informatics and Decision Making, 23(1), 192.
38. Lenyk, Z. (2021, February 3). Microsoft Vision Model ResNet-50 combines web-scale data and multi-task learning to achieve state of the art. https://www.Microsoft.Com/En-Us/Research/Blog/Microsoft-Vision-Model-Resnet-50-Combines-Web-Scale-Data-and-Multi-Task-Learning-to-Achieve-State-of-the-Art/.
39. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., & Pal, C. (2016). The Importance of Skip Connections in Biomedical Image Segmentation. https://arxiv.org/abs/1608.04117.
40. Zhang, Y., Liu, S., Li, C., & Wang, J. (2022). Application of Deep Learning Method on Ischemic Stroke Lesion Segmentation. Journal of Shanghai Jiaotong University (Science), 27(1), 99–111.
41. Wang, S.-H., Phillips, P., Sui, Y., Liu, B., Yang, M., & Cheng, H. (2018). Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling. Journal of Medical Systems, 42(5), 85.
42. Karthik, R., Gupta, U., Jha, A., Rajalakshmi, R., & Menaka, R. (2019). A deep supervised approach for ischemic lesion segmentation from multimodal MRI using Fully Convolutional Network. Applied Soft Computing, 84, 105685.

43. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 53, 197–207.
44. Karthik, R., Radhakrishnan, M., Rajalakshmi, R., & Raymann, J. (2021). Delineation of ischemic lesion from brain MRI using attention gated fully convolutional network. Biomedical Engineering Letters, 11(1), 3–13.
45. Nazari-Farsani, S., Yu, Y., Duarte Armindo, R., Lansberg, M., Liebeskind, D. S., Albers, G., Christensen, S., Levin, C. S., & Zaharchuk, G. (2023). Predicting final ischemic stroke lesions from initial diffusion-weighted images using a deep neural network. NeuroImage: Clinical, 37, 103278.
46. Yu, Y., Xie, Y., Thamm, T., Gong, E., Ouyang, J., Huang, C., Christensen, S., Marks, M. P., Lansberg, M. G., Albers, G. W., & Zaharchuk, G. (2020). Use of Deep Learning to Predict Final Ischemic Stroke Lesions From Initial Magnetic Resonance Imaging. JAMA Network Open, 3(3), e200772.
47. Shore, J., & Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on Information Theory, 26(1), 26–37.
48. Wong, K. K., Cummock, J. S., Li, G., Ghosh, R., Xu, P., Volpi, J. J., & Wong, S. T. C. (2022). Automatic Segmentation in Acute Ischemic Stroke: Prognostic Significance of Topological Stroke Volumes on Stroke Outcome. Stroke, 53(9), 2896–2905.
49. Wei, Y.-C., Huang, W.-Y., Jian, C.-Y., Hsu, C.-C. H., Hsu, C.-C., Lin, C.-P., Cheng, C.-T., Chen, Y.-L., Wei, H.-Y., & Chen, K.-F. (2022). Semantic segmentation guided detector for segmentation, classification, and lesion mapping of acute ischemic stroke in MRI images. NeuroImage: Clinical, 35, 103044.
50. Moon, H. S., Heffron, L., Mahzarnia, A., Obeng-Gyasi, B., Holbrook, M., Badea, C. T., Feng, W., & Badea, A. (2022). Automated multimodal segmentation of acute ischemic stroke lesions on clinical MR images. Magnetic Resonance Imaging, 92, 45–57.
51. Brott, T., Adams, H. P., Olinger, C. P., Marler, J. R., Barsan, W. G., Biller, J., Spilker, J., Holleran, R., Eberle, R., & Hertzberg, V. (1989). Measurements of acute cerebral infarction: a clinical examination scale. Stroke, 20(7), 864–870.
52. Winzeck, S., Hakim, A., McKinley, R., Pinto, J. A. A. D. S. R., Alves, V., Silva, C., Pisov, M., Krivov, E., Belyaev, M., Monteiro, M., Oliveira, A., Choi, Y., Paik, M. C., Kwon, Y., Lee, H., Kim, B. J., Won, J.-H., Islam, M., Ren, H., … Reyes, M. (2018). ISLES 2016 and 2017-Benchmarking Ischemic Stroke Lesion Outcome Prediction Based on Multispectral MRI. Frontiers in Neurology, 9.
53. Hernandez Petzsche, M. R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.-L., Kofler, F., Ezhov, I., Robben, D., Hutton, A., Friedrich, T., Zarth, T., Bürkle, J., Baran, T. A., Menze, B., Broocks, G., Meyer, L., … Kirschke, J. S. (2022). ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific Data, 9(1), 762.
54. Lansberg, M. G., Straka, M., Kemp, S., Mlynash, M., Wechsler, L. R., Jovin, T. G., Wilder, M. J., Lutsep, H. L., Czartoski, T. J., Bernstein, R. A., Chang, C. W., Warach, S., Fazekas, F., Inoue, M., Tipirneni, A., Hamilton, S. A., Zaharchuk, G., Marks, M. P., Bammer, R., & Albers, G. W. (2012). MRI profile and response to endovascular reperfusion after stroke (DEFUSE 2): a prospective cohort study. The Lancet Neurology, 11(10), 860–867.
55. Marks, M. P., Heit, J. J., Lansberg, M. G., Kemp, S., Christensen, S., Derdeyn, C. P., Rasmussen, P. A., Zaidat, O. O., Broderick, J. P., Yeatts, S. D., Hamilton, S., Mlynash, M., & Albers, G. W. (2018). Endovascular Treatment in the DEFUSE 3 Study. Stroke, 49(8), 2000–2003.
56. Zaharchuk, G. (2020, March 19). Imaging Collaterals in Acute Stroke (iCAS) (iCAS). Https://Clinicaltrials.Gov/Study/NCT02225730.
57. Karthik, R., Menaka, R., Johnson, A., & Anand, S. (2020). Neu2roimaging and deep learning for brain stroke detection - A review of recent advancements and future prospects. Computer Methods and Programs in Biomedicine, 197, 105728.
58. Cornelio Lea Katrina S., del Castillo, M. A. V., N. Jr. P. C. (2019). U-ISLES: Ischemic Stroke Lesion Segmentation Using U-Net. In S. and B. R. Arai Kohei and Kapoor (Ed.), Intelligent Systems and Applications. Pp. 326–336. Springer International Publishing.

59. Aboudi, F., Drissi, C., & Kraiem, T. (2022). Efficient U-Net CNN with Data Augmentation for MRI Ischemic Stroke Brain Segmentation. 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), 1, 724–728.
60. Liu, L., Kurgan, L., Wu, F., Wang, J. (2020). Attention convolutional neural network for accurate segmentation and quantification of lesions in ischemic stroke disease. Medical Image Analysis, 65:101791.
61. Ostmeier, S., Axelrod, B., Isensee, F., Bertels, J., Mlynash, M., Christensen, S., Lansberg, M. G., Albers, G. W., Sheth, R., Verhaaren, B. F. J., Mahammedi, A., Li, L.-J., Zaharchuk, G., & Heit, J. J. (2023). USE-Evaluator: Performance metrics for medical image segmentation models supervised by uncertain, small or empty reference annotations in neuroimaging. Medical Image Analysis, 90, 102927.
62. Schmidt, R. L., & Factor, R. E. (2013). Understanding sources of bias in diagnostic accuracy studies. Archives of Pathology & Laboratory Medicine, 137(4), 558–565.
63. Kim, Y.-C., Lee, J.-E., Yu, I., Song, H.-N., Baek, I.-Y., Seong, J.-K., Jeong, H.-G., Kim, B. J., Nam, H. S., Chung, J.-W., Bang, O. Y., Kim, G.-M., & Seo, W.-K. (2019). Evaluation of Diffusion Lesion Volume Measurements in Acute Ischemic Stroke Using Encoder-Decoder Convolutional Network. Stroke, 50(6), 1444–1451.
64. Abdmouleh, N., Echtioui, A., Kallel, F., & Hamida, A. ben. (2022). Modified U-Net Architecture based Ischemic Stroke Lesions Segmentation. 2022 IEEE 21st International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), 361–365.
65. Pavlakis, S. G., Hirtz, D. G., & deVeber, G. (2006). Pediatric Stroke: Opportunities and Challenges in Planning Clinical Trials. Pediatric Neurology, 34(6), 433–435.
66. Ospel, J., Singh, N., Ganesh, A., & Goyal, M. (2023). Sex and Gender Differences in Stroke and Their Practical Implications in Acute Care. Journal of Stroke, 25(1), 16–25.
67. Liu, Z., Cao, C., Ding, S., Liu, Z., Han, T., & Liu, S. (2018). Towards Clinical Diagnosis: Automated Stroke Lesion Segmentation on Multi-Spectral MR Image Using Convolutional Neural Network. IEEE Access, 6, 57006–57016.
68. Wulms, N., Redmann, L., Herpertz, C., Bonberg, N., Berger, K., Sundermann, B., & Minnerup, H. (2022). The Effect of Training Sample Size on the Prediction of White Matter Hyperintensity Volume in a Healthy Population Using BIANCA. Frontiers in Aging Neuroscience, 13.
69. Clèrigues, A., Valverde, S., Bernal, J., Freixenet, J., Oliver, A., & Lladó, X. (2020). Acute and sub-acute stroke lesion segmentation from multimodal MRI. Computer Methods and Programs in Biomedicine, 194, 105521.
70. Olivier, A., Moal, O., Moal, B., Munsch, F., Okubo, G., Sibon, I., Dousset, V., & Tourdias, T. (2019). Active learning strategy and hybrid training for infarct segmentation on diffusion MRI with a U-shaped network. Journal of Medical Imaging, 6(04), 1.
71. Zhao, B., Liu, Z., Liu, G., Cao, C., Jin, S., Wu, H., & Ding, S. (2021). Deep Learning-Based Acute Ischemic Stroke Lesion Segmentation Method on Multimodal MR Images Using a Few Fully Labeled Subjects. Computational and Mathematical Methods in Medicine, 2021, 3628179.
72. Li, C., & Ji, P. (2023). TernausNet-based segmentation of cerebral infarction in magnetic resonance images. Journal of Radiation Research and Applied Sciences, 16(3), 100619.
73. Iglovikov, V., & Shvets, A. (2018). TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. https://arxiv.org/abs/1801.05746.
74. Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255.
75. Schick, F., Pieper, C. C., Kupczyk, P., Almansour, H., Keller, G., Springer, F., Mürtz, P., Endler, C., Sprinkart, A. M., Kaufmann, S., Herrmann, J., & Attenberger, U. I. (2021). 1.5 vs 3 Tesla Magnetic Resonance Imaging. Investigative Radiology, 56(11), 680–691.
76. Cui, L., Fan, Z., Yang, Y., Liu, R., Wang, D., Feng, Y., Lu, J., & Fan, Y. (2022). Deep Learning in Ischemic Stroke Imaging Analysis: A Comprehensive Review. BioMed Research International, 2022, 2456550.
77. Wardlaw, J. M., & Farrall, A. J. (2004). Diagnosis of stroke on neuroimaging. BMJ, 328(7441), 655–656.
78. Karthik, R., & Menaka, R. (2018). Computer-aided detection and characterization of stroke lesion – a short review on the current state-of-the art methods. The Imaging Science Journal, 66(1), 1–22.

79. Khezrpour, S., Seyedarabi, H., Razavi, S. N., & Farhoudi, M. (2022). Automatic segmentation of the brain stroke lesions from MR flair scans using improved U-net framework. Biomedical Signal Processing and Control, 78, 103978.
80. Simonsen, C. Z., Madsen, M. H., Schmitz, M. L., Mikkelsen, I. K., Fisher, M., & Andersen, G. (2015). Sensitivity of Diffusion- and Perfusion-Weighted Imaging for Diagnosing Acute Ischemic Stroke Is 97.5%. Stroke, 46(1), 98–101.
81. Lee, S., Sunwoo, L., Choi, Y., Jung, J. H., Jung, S. C., & Won, J.-H. (2022). Impact of Diffusion–Perfusion Mismatch on Predicting Final Infarction Lesion Using Deep Learning. IEEE Access, 10, 97879–97887.
82. Sitburana, O., & Koroshetz, W. J. (2005). Magnetic resonance imaging: Implication in acute ischemic stroke management. Current Atherosclerosis Reports, 7(4), 305–312.
83. Kakuda, W., Lansberg, M. G., Thijs, V. N., Kemp, S. M., Bammer, R., Wechsler, L. R., Moseley, M. E., Parks, M. P., & Albers, G. W. (2008). Optimal Definition for PWI/DWI Mismatch in Acute Ischemic Stroke Patients. Journal of Cerebral Blood Flow & Metabolism, 28(5), 887–891.
84. Zhu, H., Jiang, L., Zhang, H., Luo, L., Chen, Y., & Chen, Y. (2021). An automatic machine learning approach for ischemic stroke onset time identification based on DWI and FLAIR imaging. NeuroImage: Clinical, 31, 102744.
85. Avesta, A., Hossain, S., Lin, M., Aboian, M., Krumholz, H. M., & Aneja, S. (2023). Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-Segmentation. Bioengineering, 10(2), 181.
86. Yu, L., Yang, X., Chen, H., Qin, J., & Heng, P. A. (2017). Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).
87. Wu, Z., Zhang, X., Li, F., Wang, S., Huang, L., & Li, J. (2023). W-Net: A boundary-enhanced segmentation network for stroke lesions. Expert Systems with Applications, 230, 120637.
88. Hashemi, S. R., Mohseni Salehi, S. S., Erdogmus, D., Prabhu, S. P., Warfield, S. K., & Gholipour, A. (2019). Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection. IEEE Access, 7, 1721–1735.
89. Zhang, Y., Liu, S., Li, C., & Wang, J. (2021). Rethinking the Dice Loss for Deep Learning Lesion Segmentation in Medical Images. Journal of Shanghai Jiaotong University (Science), 26(1), 93–102.
90. Hu, X., Luo, W., Hu, J., Guo, S., Huang, W., Scott, M. R., Wiest, R., Dahlweid, M., & Reyes, M. (2020). Brain SegNet: 3D local refinement network for brain lesion segmentation. BMC Medical Imaging, 20(1), 17.
91. Rachmadi, M. F., Poon, C., & Skibbe, H. (2023). Improving Segmentation of Objects with Varying Sizes in Biomedical Images using Instance-wise and Center-of-Instance Segmentation Loss Function. https://arxiv.org/abs/2304.06229.
92. Inamdar, M. A., Raghavendra, U., Gudigar, A., Chakole, Y., Hegde, A., Menon, G. R., Barua, P., Palmer, E. E., Cheong, K. H., Chan, W. Y., Ciaccio, E. J., & Acharya, U. R. (2021). A Review on Computer Aided Diagnosis of Acute Brain Stroke. Sensors, 21(24).
93. Wang, S., Tan, S., Gao, Y., Liu, Q., Ying, L., Xiao, T., Liu, Y., Liu, X., Zheng, H., & Liang, D. (2018). Learning Joint-Sparse Codes for Calibration-Free Parallel MR Imaging. IEEE Transactions on Medical Imaging, 37(1), 251–261.
94. Babu, M. S., & Vijayalakshmi, V. (2019). A review on acute/sub-acute ischemic stroke lesion segmentation and registration challenges. Multimedia Tools and Applications, 78(2), 2481–2506.
95. Liu, L., Wu, F.-X., & Wang, J. (2019). Efficient multi-kernel DCNN with pixel dropout for stroke MRI segmentation. Neurocomputing, 350, 117–127.
96. Choi, Y., Kwon, Y., Lee, H., Kim, B. J., Paik, M. C., & Won, J.-H. (2016). Ensemble of Deep Convolutional Neural Networks for Prognosis of Ischemic Stroke. In A. Crimi, B. Menze, O. Maier, M. Reyes, S. Winzeck, & H. Handels (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Pp. 231–243). Springer International Publishing.
97. Pinto A., Pereira, S., M. R., A. V., W. R., S. C. A., R. M. (2018). Enhancing Clinical MRI Perfusion Maps with Data-Driven Maps of Complementary Nature for Lesion Outcome Prediction. In J. A. and D. C. and A.-L. C. and F. G. Frangi Alejandro F. and Schnabel (Ed.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Pp. 107–115. Springer International Publishing.

98. Duan W., Zhang, L. and C. J. and G. G. and Y. X. (2021). Multi-modal Brain Segmentation Using Hyper-Fused Convolutional Neural Network. In S. M. and H. M. and K. V. and R. J. M. and T. C. and W. T. Abdulkadir Ahmed and Kia (Ed.), Machine Learning in Clinical Neuroimaging. Pp. 82–91. Springer International Publishing.
99. Zhang, L., Song, R., Wang, Y., Zhu, C., Liu, J., Yang, J., & Liu, L. (2020). Ischemic Stroke Lesion Segmentation Using Multi-Plane Information Fusion. IEEE Access, 8, 45715–45725.
100. Lucas, C., Kemmling, A., Mamlouk, A. M., & Heinrich, M. P. (2018). Multi-scale neural network for automatic segmentation of ischemic strokes on acute perfusion images. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 1118–1121.
101. Karthik, R., Menaka, R., Hariharan, M., & Won, D. (2021). Ischemic Lesion Segmentation using Ensemble of Multi-Scale Region Aligned CNN. Computer Methods and Programs in Biomedicine, 200, 105831.
102. Vupputuri, A., Gupta, A., & Ghosh, N. (2021). MCA-DN: Multi-path convolution leveraged attention deep network for salvageable tissue detection in ischemic stroke from multi-parametric MRI. Computers in Biology and Medicine, 136, 104724.
103. Ou, Y., Yuan, Y., Huang, X., Wong, K., Volpi, J., Wang, J. Z., & Wong, S. T. C. (2021). LambdaUNet: 2.5D Stroke Lesion Segmentation of Diffusion-Weighted MR Images. Pp. 731–741.
104. Cao, Y., Liu, W., Zhang, S., Xu, L., Zhu, B., Cui, H., Geng, N., Han, H., & Greenwald, S. E. (2022). Detection and Localization of Myocardial Infarction Based on Multi-Scale ResNet and Attention Mechanism. Frontiers in Physiology, 13.
105. Liu, C., Yin, Y., Sun, Y., & Ersoy, O. K. (2022). Multi-scale ResNet and BiGRU automatic sleep staging based on attention mechanism. PLOS ONE, 17(6), e0269500.
106. Marcos, L., Quint, F., Babyn, P., & Alirezaie, J. (2022). Dilated Convolution ResNet with Boosting Attention Modules and Combined Loss Functions for LDCT Image Denoising. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 1548–1551.
107. Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? https://arxiv.org/abs/1703.04977.
108. Richardson, M., Garner, P., & Donegan, S. (2019). Interpretation of subgroup analyses in systematic reviews: A tutorial. Clinical Epidemiology and Global Health, 7(2), 192–198.
109. Hedges, L. (1985). Statistical Methods for Meta-Analysis. Elsevier.
110. Kumar, A., Upadhyay, N., Ghosal, P., Chowdhury, T., Das, D., Mukherjee, A., & Nandi, D. (2020). CSNet: A new DeepNet framework for ischemic stroke lesion segmentation. Computer Methods and Programs in Biomedicine, 193, 105524.
111. Karpathy, A., & Fei-Fei, L. (2017). Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), 664–676.
112. Pellegrini, E., Ballerini, L., Hernandez, M. del C. V., Chappell, F. M., González-Castro, V., Anblagan, D., Danso, S., Muñoz-Maniega, S., Job, D., Pernet, C., Mair, G., MacGillivray, T. J., Trucco, E., & Wardlaw, J. M. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10(1), 519–535.
113. Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for Simplicity: The All Convolutional Net. https://arxiv.org/abs/1412.6806.
114. Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. Pp. 818–833.
115. Qin, S., Zhang, Z., Jiang, Y., Cui, S., Cheng, S., & Li, Z. (2023). NG-NAS: Node growth neural architecture search for 3D medical image segmentation. Computerized Medical Imaging and Graphics, 108, 102268.
116. Allcock, J., Vangone, A., Meyder, A., Adaszewski, S., Strahm, M., Hsieh, C.-Y., & Zhang, S. (2022). The Prospects of Monte Carlo Antibody Loop Modelling on a Fault-Tolerant Quantum Computer. Frontiers in Drug Discovery, 2.