# **Title: Deep Learning for MRI-based Acute and/or Subacute Ischemic Stroke Lesion Segmentation - A Systematic Review and Pilot Analysis.**

**Name**: Makram Baaklini (M.B.)

**Student number**: s2264749

**Submission date**: April 21st, 2024

**Supervisor**: Dr. Maria Valdes Hernandez (M.V.H.)

# Declaration

I, Makram BAAKLINI, declare that the project report is my own and original work and has not previously been submitted for any other assessment for the purposes of this or any other award.

# Abstract

## Background

We systematically reviewed the literature (2015-2023) for deep learning algorithms that segment acute stroke lesions on brain MRI. We compared their accuracy and generalizability, while providing tips to developers.

## Methods and Materials

We registered the systematic review protocol with PROSPERO (ID:CRD42023481551). We completed our search in IEE Explore, MEDLINE, ScienceDirect, Web of Science, PubMed, Springer, and OpenReview.net. We assessed the risk of bias of retained studies and their methodological quality using NIH’s study quality assessment tool[²](#Footnote2). We conducted an in-depth meta-analysis of our findings, and proposed a stroke lesion segmentation algorithm, “AG-UResNet50”, in light of the review’s key findings.

## Results

The search yielded 1485 papers, of which 39 were ultimately retained. 13/39 studies incorporated attention mechanisms in their architecture, and algorithm performance was measured using “Dice” in 37/39 studies.

## Conclusion

We found no evidence that favours using attention mechanisms in deep learning architectures for stroke lesion segmentation on MRI data. Also, the generalizability of the proposed algorithms was generally below par.

\***Abbreviations are listed at the end of the document** (cf.[Abbreviations](#Abbreviations)).

# Introduction

Stroke remains a leading cause of mortality and long-term disability worldwide [14], placing a substantial burden on healthcare systems and societies [15]. The majority of strokes are ischemic, and they can occur in different locations and various shapes [18]. Ischemic penumbra denotes an “at risk” region that is functionally impaired, but potentially salvageable, in contrast with the ischemic core which refers to the irreversibly damaged tissue [19].

After stroke onset, the progression of ischemic injury continues for minutes-to-days, depending on brain region vulnerability, cellular constituents, and residual perfusion levels [20]. Accurate diagnosis during acute-to-subacute stages allows for interventions (e.g., thrombolytic drugs, surgery) that may potentially salvage the penumbral area.

Today, MRI technology enables the non-invasive investigation of human brain features, thanks to the high dimensionality and particularly low signal-to-noise ratio found in MRI scans.

Since manual segmentation methods are time-consuming and subject to inter-rater variability, there has been a growing interest, since 2015 [9], in applying DL techniques to automate stroke lesion segmentation tasks and enhance their accuracy. DL methods can automatically extract intricate spatial and textural features within MRI scans, while requiring low-to-moderate subject matter expertise.

DL also addresses long-dated ML-related challenges, such as the curse of dimensionality, which occurs when the number of dimensions in the data increases, as it is the case with imaging data.

In this systematic review, we will address the following question: “**What would be the optimal deep learning model architecture for acute and/or subacute ischemic stroke lesion segmentation on brain MRI?**”

We investigated the accuracy and generalizability of the proposed DL methods in acute-to-subacute stroke lesion segmentation on MRI, through a systematic review and meta-analysis, as well as a pilot analysis aiming to address as many of the limitations identified in the review as possible.

While several related reviews have been published since 2017 [1-12,16,73,82,83], their scopes differ from ours ([cf.Fig.1](#Fig1)). Our review is particularly robust, as it supplements the meta-analysis by a pilot analysis, and it focuses on DL architectures and attention mechanisms in much higher granularity than existing reviews.

# Background

## Deep Learning Architectures

CNNs are useful for processing data with grid-like topology (e.g., 2D/3D grid of pixels/voxels) [32]. They use convolution blocks to produce “feature maps”. CNNs use sparse inter-layer interactions by making the kernel smaller than the input [33].

A standard convolution block in a CNN ([cf.Fig.2a](#Fig2)) consists of a linear convolution operation on a 3x3 kernel, which produces a feature map that is passed through a ReLU activation function to introduce non-linearity and enable the network to learn more complex relationships in the data [10], before it gets down-sampled by a pooling operation.

CNNs are widely used in medical image segmentation [54], with an architecture that typically ends with fully-connected layer(s) responsible for doing predictions (e.g., pixel/tissue classification). Predictions are connected to a 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 [10].

CNNs’ downsides are that (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 by (i) replacing CNN's fully-connected layer(s) by “up-sampling convolutions” that output images of the same size as the input, and (ii) generating likelihood maps instead of pixel-by-pixel predictions. However, FCN’s output maps are of particularly low resolution [9].

U-Net architecture was first used for image segmentation in 2015 [22], and it has since achieved overwhelming success. It uses a symmetric encoder-decoder structure based on convolution blocks, where down-sampling (encoder) operations compress images and up-sampling (decoder) operations restore them, until they reach the input image’s original size [52], 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** [5].

ResNet architecture was published shortly after U-Net [84], to further tackle the vanishing gradient problem, also using skip connections. A standard ResNet block ([cf.Fig.2b](#Fig2)) consists of an “identity path” (green arrow) 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 [31].

ResNets can improve model convergence speed [30], but since most residual blocks only slightly change the input signal, they produce a large amount of redundant features [27]. This is where DenseNets help.

DenseNet architecture was published shortly after ResNet [61]. It employs dense connections interconnecting all layers, in order to maximize information and gradient propagation [5].

A standard Dense block is represented in [Fig.2c](#Fig2). 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 [52]. Reusing features across layers also allows DenseNets to tackle the vanishing gradient problem [28].

## 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 [48]. 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 [40]. In this paper, we categorize attention mechanisms as “spatial”, “channel”, or “hybrid”.

“Spatial attention” (cf.[Fig.3a](#Fig3)) 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 [48].

Examples of spatial attention methods include:

* Attention gates (cf.[Fig.5](#Fig5))
* Self-attention: it operates solely on input sequences, thus enabling a model to exploit spatial relationships within input scans [51]
* Cross-attention (e.g., Gomez et al. [79]): it 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 decrease noisy signals in skip connections [51]

"Channels" represent individual feature maps stacked in a tensor. Each map specializes in detecting specific features (e.g., horizontal edges, brain anatomy). “Channel attention” (cf.[Fig.3b](#Fig3)) refers to the process of **assigning a weight to each feature map**, emphasizing those that contribute most significantly to the learning [48]. Conversely, spatial attention **assigns weights to** **pixels**.

Examples of channel attention methods include squeeze-and-excitation blocks [49], which were used by Woo et al. [43] and Lee A. et al. [46]).

**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: it combines spatial and channel attention gates (sAG+cAG) [80]
* Multi-head attention: it uses parallel processing by applying attention across multiple "heads" simultaneously, where each head may be configured to implement different attention types, including channel and spatial attention [51]

# Methods and Materials

## Systematic Review & Meta-analysis

### Literature Search

We performed the review following the NIH’s Study Quality Assessment Tool[²](#Footnote2) and PRISMA guidelines[5](#Footnote5).

We conducted a literature search (2015-2023) for papers published in PubMed, MEDLINE, IEEE Xplore, Web of Science, ScienceDirect, and Springer. We identified keywords by expanding 5 subject components: Accuracy, Acute ischemic stroke, Deep learning, Lesion segmentation, MRI.

We also did citation tracking of reviewed articles, and hand-searching of 2 journals, “Stroke” and “NeuroImage: Clinical” (Recall:100%). 2 reviewers (M.B and M.V.H.) conducted the main search, paper selection, and data extraction. The full search strategy can be found in [Appendix.A.A](#AppendixA).

### Eligibility Criteria

cf.[Table.1](#Table1).

### Data Extraction

For each paper, we extracted the following information

* Basic information (e.g., DOI, first author)
* 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
* Generalizability result (custom calculation: cf.[Appendix.A.C](#AppendixA))
* NIH Study Quality Assessment (cf.[Appendix.A.C](#AppendixA))

We thoroughly analysed the obtained results using python. All visuals created are in [Appendix.A.B](#AppendixA).

### Whole Group Analysis

cf.[Table.2](#Table2).

### Subgroup Analysis

cf.[Table.3](#Table3).

### Sensitivity Analysis

cf.[Table.4](#Table4).

### Meta-regression

cf.[Table.5](#Table5).

### Publication Bias

cf.[Table.6](#Table6).

### Methodological Quality Assessment

Most DL-based stroke lesion segmentation studies use existing datasets (e.g., ISLES), so they are “observational”, since they do not involve new clinical trials or modifications to patient populations.

Accordingly, we used NIH’s QA tool for observational studies[²](#Footnote2) for the methodological QA of retained studies.

## Pilot Analysis

The pilot analysis consisted of the deployment of a DL algorithm, “AG-UResNet50”, for AIS lesion segmentation on brain MRI. The aim of this pilot is two-folds:

* **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
  + Compound versus region-based loss functions
  + Single-modality versus multi-modality approach

### Dataset

We used the ISLES-2015-SISS dataset, published by the MICCAI 2015 conference [13]. It consists of 28 subacute stroke cases to use for model training. For each case, a set of 5 sequences was provided: T1-WI, T2-WI, DWI, and FLAIR, along with the corresponding ground-truth masks.

The data was already anonymized by removing patient information from files and facial bone structure from images.

### Data Pre-processing

The following data pre-processing activities 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)

### Proposed Method

Inspired by multiple papers [86-90,94], especially Guerrero et al.’s UResNet [89], Jin et al.’s RA-UNet [94], and Gheibi et al.’s CNN-Res [85], “AG-UResNet50” is a 5-level end-to-end U-Net (cf.[Fig.4](#Fig4)), with as main novelty the replacement of its encoder path by ResNet50 [78].

Using U-Net in combination with ResNet50 allows us to leverage the power of skip connections further [91], and to 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. Moreover, Zhang et al. [73] identified ResNet as an architecture that can improve segmentation of small lesions.

Max-pooling was only 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 2 were used for remaining down-sampling operations, in order to better retain image details [5].

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 [92], as also demonstrated by Karthik et al. (1) [23]. 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 [40], as seen in Karthik R. et al. (2) [34], Nazari-Farsani et al. [38], and Yu et al. [39].

An AG takes 2 input vectors that are added element-wise (cf.[Fig.5](#Fig5)), 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 [59], while Dice loss directly computed the gradient using the Dice score of predicted versus real samples [61].

From a regularization standpoint, pixel dropout, learning rate adjustment and data augmentation methods were used, while for optimization, Adam function and batch normalization were used.

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.

### Model Evaluation

Dice Similarity Coefficient (DSC) was used to evaluate model performance. It measures the spatial overlap between 2 segmentations A and B (e.g., predictions and masks):

* DSC(A,B)=2(A∩B)/(A+B), where ∩=intersection.

We used 5-fold cross-validation for model validation.

# Results

## Systematic Review & Meta-analysis

### Search Results

The search yielded 1485 papers, of which 39 were ultimately retained (cf.[Fig.6](#Fig6)).

The main reasons for discarding papers were the following (ordered by relevance):

* ML-based rather than DL
* CT-based
* Inclusion of hyper-acute and/or chronic stroke cases (e.g., ATLAS2.0-based studies)
* Classification algorithm running after the segmentation **AND** intermediary segmentation results not reported (e.g., some DeepMedic-based studies)
* Segmentation results only provided on training set
* Focus on haemorrhagic stroke
* Model architecture not explained
* Focus on brain tumour and/or WMH lesions
* Segmentation results reported specifically for core and penumbra (not overall performance)
* Using synthetic data
* Not accessible via institutional login

### Data Extraction Results

Clearly, all papers had “segmentation” as primary outcome. Fewer had “prognosis” (6 studies) or “functional” (3 studies) outcomes. For prognosis, studies were either trying to predict tissue fate or lesion volume (e.g., Wong et al. [24], Wei et al. [29]), while for functional, they mostly tried to predict the modified ranking scale score (mRS). While only 1 paper had “diagnosis” as outcome, diagnosis is tightly linked to segmentation, as by identifying lesion pixels, we are effectively helping physicians perform their diagnosis of stroke.

#### Sample Characteristics

In terms of patient populations (cf.[Table.7](#Table7)), patients were all at least 18 y.o., and males were generally over-represented (Mean:58%) (cf.[Fig.7a](#Fig7)), except in a few studies where the opposite was true (e.g., Moon et al. [65]).

From a stroke severity standpoint, reported mean NIHSS [45] were always on the “minor” or “moderate” ranges (7 studies) (cf.[Fig.7d](#Fig7)).

Although both subacute and acute stroke stages were in scope, most studies (23/39) included exclusively AIS cases. Reported patient mean “time-since-stroke” (TSS) were also exclusively in the acute interval (cf.[Fig.7b](#Fig7)), with 2 studies actually very close to the hyperacute-acute border.

Only 4 papers used sample sizes above 500 (Mean:252.2) (cf.[Fig.7c](#Fig7)), and samples were most often collected from 1 center (28 studies versus 14 from multiple centers).

#### Medical Imaging

Most studies (25/39) used images of high or very high spatial resolution, and logically, very high resolution images were used when mean lesion volumes were relatively small (<35ml) (cf.[Fig.10c](#Fig10)).

DWI modality was by far the most used (37 studies), followed by FLAIR (17 studies). Also, two-thirds of studies adopted a multi-modal approach.

27 studies were 2D-based and 10 3D-based (cf.[Table.8](#Table8)). 2D models exclusively used high or very high-resolution images, and 3D models exclusively moderate or low-resolution images, which seems counter intuitive (cf.[Fig.8a](#Fig8)). Logically, 3D models adopted patch-wise training in 8/10 studies (cf.[Fig.8b](#Fig8)).

During the labelling of ground-truth images, 13 studies leveraged DWI-PWI mismatch, and 12 DWI-FLAIR mismatch.

The magnetic field of the scanner(s) was for 25 studies both 1.5T & 3T, for 9 studies only 3T, and for 3 only 1.5T.

#### Data Pre-processing

18 studies used proprietary datasets (cf.[Table.9](#Table9)), 24 used one or a combination of ISLES-2015 [13], ISLES-2017 [97] or ISLES-2022 [21], and 3 used data related to the DEFUSE or iCAS studies [98-100].

In 37 cases, intensity-based skull-stripping was performed (using BET2/ITK software), once atlas-based (Moon et al. [65] using Kirby/MMRR template), and once DL-based, to reduce sensitivity to noise [83] (Liu C. et al. [80] using in-house "UNet BrainMask"). Interpatient image registration onto a standard space (e.g., MINI) and intra-patient registration (e.g., registration of different sequences) was performed in 27 studies.

#### Deep Learning Architectures

Dice loss was used in 23 papers (cf.[Table.10](#Table10)), either mixed with other loss functions (8 studies) or on standalone (15 studies). Cross-entropy loss was used in 17 papers, 9 times on standalone. Focal loss was only used in 4 papers, and 2 papers used Liu et al.’s custom-built loss function [101].

37/39 studies did semantic lesion segmentation, which is tied to the fact the 35 studies deployed U-Net-based models (cf.[Fig.9](#Fig9)), and U-Net is a semantic segmentation algorithm. None of the studies have used the original U-Net as-is [22], with perhaps Cornelio et al. [41] and Aboudi et al. [55] staying the closest. ResNet architecture was the second most used (8 studies), while DenseNets were only used in 3 studies.

Data augmentation was the most used regularization method (28 studies), whereas each of dropout, ES, weight decay, class weighting, and learning rate adjustment was used in ~10 studies.

More papers used image-wise training (27 studies versus 14 for patch-wise training), but patch-wise training was used in 7/8 studies that were dealing with smaller mean lesion volumes (<40ml).

In addition, none of the papers performed uncertainty quantification, and 31 algorithms were end-to-end (versus 8 multi-module).

#### Attention Mechanisms

12 studies used attention (cf.[Table.11](#Table11)): 5 used hybrid attention, 4 spatial attention, and 3 channel attention.

Oddly enough, papers deploying ResNet-based architectures never incorporated attention.

#### Performance and Generalizability

The most used performance metrics across studies were Dice, Recall, Precision (overlap metrics), and HD (distance metric) [58] cf.[Table.12](#Table12). 6 papers only reported 1 metric.

We saw a clear positive correlation between (i) Dice and Precision scores (cf.[Fig.10b](#Fig10)), and (ii) Generalizability and Dice scores (cf.[Fig.10a](#Fig10)).

Generalizability scores were manually calculated based on sample representativeness, ground-truth data, and access to clean code (cf.[Appendix.A.C](#AppendixA)). Only Liu C. et al.’s algorithm [80] was deemed “highly” generalizable, whereas 19 algorithms had “low” generalizability.

Only 7 papers have reported segmentation performance on small versus large lesions, and in 5 of them, accuracy on small lesions was lower or significantly lower. We also saw a positive correlation between spatial resolution and segmentation performance (cf.[Fig.10d](#Fig10)).

Only 7 studies have performed external validation of their models on unseen data, most of which got higher Dice on the study’s test set than on the external validation set.

There was positive correlation between sample size and segmentation performance, and single-center studies showed better performance than multi-center studies (Mean Dice:0.71 versus 0.6).

Lesion volume ranges were substantially different 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 scores (>0.68). In other words, segmentation performance was generally better when lesions were larger.

Dice scores were much higher for studies using ISLES-2015 (Mean Dice:>0.7) or proprietary (Mean Dice:0.72) datasets, than when using ISLES-2017 (Mean Dice:0.38) or DEFUSE (Mean Dice:0.52).

When attention-based networks were deeper (cf.[Fig.10f](#Fig10)), or when U-Nets were deeper (cf.[Fig.10e](#Fig10)), Dice scores were higher. 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 1 of the algorithms that used a relatively high number of epochs was also using ES regularization, which means that for all the others, the full (high) amount of epochs was used during training, hence increasing the probability of overfitting substantially.

### Risk of Bias Assessment

We assessed the possibility that the proposed algorithms did not fully measure what they intended to measure due to biases, which would in turn make them less “valid”.

31 studies scored “GOOD”, and 8 scored “FAIR” in the NIH study QA[²](#Footnote2). Although these results are positive, we have identified several 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)
* Age range of participants not always consistent (e.g., Kim et al. [60] only included patients between 58-79 y.o.)

We also noticed cases of **selection bias**:

* Multiple studies used the same ISLES datasets to evaluate the performance of their segmentation methods, and that, although advantageous (e.g., cost effective, allows comparability) introduces selection bias
* Males were generally over-represented

Also, ground-truth data was most often obtained by manually refining semi-automatic segmentations (e.g., thresholding followed by region-growing), which introduces **observer bias**. 16 studies did not provide information about labelling criteria, so it is unclear whether observer bias was present in those.

We noted 2 other forms of bias:

* **Verification bias**: In 10 studies, only 1 expert did the labelling of ground-truth images
* **Measurement bias**:
  + Mean Dices on ISLES-2017 were generally much lower than those on ISLES-2015
  + 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

### Whole Group Analysis

We would like to further assesses the segmentation performance of the proposed algorithms, as reflected by Dice scores (effect size). 17 papers have reported their Dice along with STD, so only 17 studies were included in this analysis, since STD was needed to compute the inverse variance (effects method).

The forest plot (cf.[Fig.11](#Fig11)) shows per study, the Dice, STD, sample size, and weight assigned from both the fixed-effects (FE) and random-effects (RE) analyses.

We performed heterogeneity tests using both FE and RE models:

* **I²**=22.08%
* **FE**: Heterogeneity: Q=0.05(p=1.00); Overall effect test: z=17.78(p=0.00)
* **RE**: Heterogeneity: Q=21.82(p=0.15); Overall effect test: z=2.06(p=0.04)

I² describes the percentage of variation across studies that is due to heterogeneity rather than chance [78]. An I² of 22.08% indicates a low-to-moderate level of heterogeneity between studies.

FE assumes there is one true Dice for all studies, and any differences between observed Dice scores are due to random fluctuations or sampling error (H0):

* Q-stat was disregarded since p=1.0>0.05. Besides, Q was shown to have low power as a test of heterogeneity, when the number of studies is relatively small [77]
* The very high Z-stat (17.78,p=0.0) indicates a substantially large effect size (i.e., ~17 STDs away from the mean in either direction)

Given the lack of evidence for heterogeneity and the highly significant Z-stat, we cannot confidently accept/reject H0 under the FE model. In order to account for potential heterogeneity, and to provide a more conservative estimate of overall effect size, we will conduct a RE analysis.

The RE model tests the null hypothesis (H0) that there is more variability in Dice scores between studies than expected by chance alone:

* Q-stat was higher (21.82), suggesting that there might be some heterogeneity, but the evidence is not robust enough (p=0.15>0.05). However, this is aligned with the value obtained for I²
* Z-stat (2.06,p=0.04) indicates a statistically significant overall effect

Given the observed potential heterogeneity in the RE analysis, we can conclude that the **RE model allows for a more conservative estimation of variability in Dice scores between retained studies**.

### Subgroup Analysis

In order to further evaluate the association between “attention mechanisms” and “Dice scores”, we performed a subgroup analysis. The resulting forest plot is shown in [Fig.12](#Fig12).

We then performed heterogeneity tests on each subgroup, using RE model:

* **Subgroup “with attention”**: Heterogeneity: Q=7.31(p=0.12), I²=31.63%; Overall effect test: z=39.03(p=0.0)
* **Subgroup “without attention”**: Heterogeneity: Q=19.93(p=0.05), I²=39.8%; Overall effect test: z=5.27(p=0.0)

Subgroup “with attention” indicated moderate heterogeneity in I² (31.63%) and a very high Z-stat (39.03,p=0.0), suggesting a substantially large overall effect. While this suggests that the presence of attention may enhance segmentation performance, the small number of studies in this subgroup (5) limits the conclusiveness of this finding.

In contrast, subgroup “without attention” comprised 12 studies, showing significant heterogeneity in the Q-stat (Q=19.93,p=0.05) and in I² (39.80%). Despite the absence of attention, a significant overall effect was also observed (z=5.27,p≈0). This suggests that **even when not using attention, there are significant differences in Dice scores between studies**.

We then conducted the **two-tailed subgroup differences test**:

* Difference in overall effect sizes: 0.06
* Standard error (SE) of difference: 0.04
* z=1.39(p=0.17)

Subgroup “with attention” showed slightly larger effect size (+0.06) (i.e., slightly higher Dices), but the Z-stat is not statistically significant (p=0.17>0.05). Therefore, we cannot reject H0, although +0.06 does appear to be a **non-significant difference in overall effect sizes between these subgroups**.

### Sensitivity Analysis

We performed a sensitivity analysis by using Precision scores as effects size instead of Dice scores. Only 8 studies have reported their precision scores along with STD (cf.[Fig.13](#Fig13)).

When re-assessing heterogeneity with this modified effect size, we obtained:

* **I²**=8.49%
* **FE**: Heterogeneity: Q=0.07(p=1.00); Overall effect test: z=6.4(p=0.00)
* **RE**: Heterogeneity: Q=8.74(p=0.27); Overall effect test: z=1.47(p=0.14)

I² decreased from 22.08% to 8.49% when using Precision as effect size measure, indicating a reduced level of heterogeneity between studies compared to the Dice-based analysis.

In FE, Q-stat was disregarded again (p=1.0>0.05), and in RE, the evidence provided by the Z-stat (1.47) was not statistically significant (p=0.14>0.05).

**This sensitivity analysis is therefore** **not conclusive**, as we were unable to properly assess whether using Precision instead of Dice as effect size measure provides a robust heterogeneity assessment.

### Meta-regression

The aim of this meta-regression was to further assess the statistical significance of the relationship between “attention mechanisms” and “Dice scores” (cf.[Fig.14](#Fig14)).

8.1% of the variance in Dice scores is explained by the presence of attention (R-squared: 0.081).

The slope indicating the change in Dice associated with the presence of attention is not statistically significant (0.117,p=0.27>0.05), so we cannot conclude that the presence of attention has a significant impact on the likelihood of high Dice.

Meanwhile, the 95% CI of the slope [-0.100,0.334] includes 0, suggesting that **the effect of attention on Dice is not statistically significant at p=0.05**.

### Publication Bias

The assessment of publication bias was based on reported Mean Dice scores.

The funnel plot is asymmetric (cf.[Fig.15](#Fig15)), suggesting **presence of publication bias**. The high concentration of data points on the right side of the green line (i.e., weighted mean of reported mean Dice scores) (11/17 studies) suggests selective reporting of studies with high Dice. Specifically, the 6 points with very low SE (upper-right corner) represent the studies that were the most likely to be published thanks to their favourable results.

Conversely, the 5 data points with medium SE on the left side of the plot, just outside the lower 95% CI bound, indicate studies with effect estimates considerably lower than the overall weighted mean Dice.

The 6 data points inside the 95% CI bounds (within the red diagonals) refer to studies whose mean Dice scores were consistent with overall findings.

To further investigate this asymmetry, we conducted **Egger's test** (cf.[Fig.16](#Fig16)). The slope indicates the change in SE associated with a one-unit increase in Dice. Since it is negative (-0.033), it also suggests that higher Dice is associated with lower SE, with statistical significance (p=0<0.05), **thus confirming the previously observed publication bias**.

## 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 (cf.[Table.13](#Table13)).

The second best was “AG-UResNet50” (0.676±0.222), with a single-modality approach, and using the same compound loss.

Some other interesting results included:

* Experiments with “UNet” and “AG-UNet” have generated relatively poor Dice scores
* Performance was better in single-modality experiments. Abdmouleh et al. [63] 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 better when using compound loss “BCE=0.3+Dice=0.7” versus the other 2
* The 12 experiments using attention and the 12 not using attention got quasi-equal average Dice scores

Average training times are reported in [Table.14](#Table14). Multi-modal experiments took longer to train in all cases (~3h longer each time). Same was true for attention-based experiments (~30min longer each time).

The full code used for this pilot has been published in GitHub[1](#Footnote1).

# Discussion

## Systematic Review & Meta-analysis

Although our review was not restricted to adults, samples never included patients below 18 y.o. This stresses the lack of research in paediatric stroke, which may be due to multiple factors (e.g., delayed identification of stroke, numerous stroke etiologies and risk factors in children, limited imaging data‌ [64]).

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), or more severe pre-stroke disability among females [66].

It makes sense that studies focused exclusively on minor-to-moderate stroke cases, as it is complicated for severe stroke cases to undergo MRI scans (e.g., need to remain still for a prolonged period, impaired decision-making). The focus on acute stroke Is also understandable, since DWI and FLAIR are able to show high signal in AIS-affected brain areas, whereas signal begins to gradually diminish as of the subacute stage, often leading to lower MRI sensitivity for stroke [21].

Actually, such differences in MRI signals between subacute lesions and acute lesions give the idea that combining acute and subacute cases in 1 single dataset, as seen in Liu C. et al. [80] and Liu Z. et al. [118], might more easily introduce errors in the manually labelled data.

We also noticed relatively small sample sizes across studies, which is not new in AIS research [105]. Data augmentation is a common way to mitigate this issue, and Clèrigues et al. [75] 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. [17]), semi-supervised learning using weakly labelled data (e.g., Zhao et al. [109]), or transfer learning (e.g., Li et al. [67] used TernausNet [68] which was pre-trained on ImageNet [70]).

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 [117].

Most studies used DWI, known as the gold standard for early stroke detection [2], and many used T1-WI, a staple in subacute stroke research [111], T2-WI, PWI, or FLAIR. PWI is frequently applied to detect ischemic penumbra [83], and FLAIR offers enhanced lesion clarity by suppressing CSF details [6]. For instance, Khezrpour et al.‘s U-Net used only FLAIR and got very high accuracy [96]. 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.

DWI-PWI mismatch [122] was commonly used to create ground-truth sets (e.g., Lee S. et al. [25]), since PWI identifies penumbral tissue, while DWI delineates the core infarct (i.e., areas of restricted water diffusion [111]). Despite its utility though, DWI-PWI mismatch analysis remains challenging:

* Establishing clear imaging boundaries for recoverable tissue is not straight-forward [111]
* Large perfusion abnormalities may be observed in patients without corresponding clinical deficits [108]
* There is no universally defined mismatch ratio, although Kakuda et al. tried to define one [93]

DWI-FLAIR mismatch, on the other hand, is mostly used for TSS assessment in hyper-acute-to-early-acute stage [120].

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 [115], 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 [54].

Cutting 3D images into 3D patches (i.e., patch-wise training) is a way to mitigate both the computational challenges, by reducing memory overhead [5], and the small lesions challenge, by forcing the model to focus on a smaller area of the entire image. That explains why 8/10 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 [56]. For instance, Zhang R. et al. [28] 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 [110], 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 generalizability.

Moving on to loss functions, CE loss quantifies the difference between 2 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 [81].

Focal loss is an adaptation of CE loss that introduces a modulating factor aimed at down-weighting the impact of well-classified examples [57], 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 [62]).

Generally, overlap-based loss functions (e.g., Dice loss) are more robust to data imbalance issues [57]. 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 [57]:

* It fails to capture the distance between non-overlapping but close lesions
* It overlooks precise contour details. Combining it with a boundary-based loss may help
* It disproportionately penalizes small lesions, especially in presence of large lesions, as opposed to distribution-based loss (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” [69], loss with data fusion [4]).

Since most studies were U-Net-based, they primarily performed semantic lesion segmentation. Perhaps the fact that only 2 studies did instance segmentation is linked to the difficulty of delineating individual lesions in presence of motion artefacts and irregular shapes [42,44], as shown by Wu et al. [110].

Meanwhile, several studies proposed quite innovative methods:

* Liu Z. et al. [118] proposed a ResNet and a global convolution network-based (GCN) encoder-decoder. Each modality was concatenated to a 3-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
* Liu L. et al. (3)‘s MK-DCNN [116] consisted of 2 sub-DenseNets with different convolution kernels, aiming to extract more image features than with a single kernel by combining low and high resolution
* 3 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 [5] and/or to decrease prediction variance (e.g., Choi et al. [26])
* Wu et al.’s W-Net [110] 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. [47], Duan et al. [53] and Zhang L. et al. [50] 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. [71] 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. (3) [114] 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. [113] 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. [112] 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 retained studies, as numerous relevant publications combine ResNet with attention [123-125].

In terms of optimization methods, RMSProp can be effective in DL (e.g., Ou et al. [112]), as it is able to discard history from 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 [95] 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 [9].

The generalizability 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 easily improve the generalizability 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, but several cases of potential or real bias, were caused by "lacks" (e.g., lack of variability in experiments, lack of follow-up scans, lack of clear definitions of study populations and inclusion criteria). Here is a summary of our conclusions:

* 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 ISLES datasets may limit the generalizability of results
* Findings in terms of segmentation of small versus large lesions are slightly flawed, due to the various ways in which these 2 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 [107]

Our whole group analysis included 17 studies, which is enough to consider findings meaningful [121]. 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 [106].

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, etiology), 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).

## 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, although we applied data augmentation and pixel dropout to mitigate this issue.

Not using attention proved to be slightly better than using attention. Arguably, in this specific implementation where the sample was small and the network already relatively deep, **increasing the number of learnable parameters using attention gates might have accentuated the overfitting problem**.

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 (2) [30] and Liu Z. et al. [118], who did the same comparison of 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 [83].

Compound loss (Dice+CE) outperformed Dice loss, as it was the case with Kumar et al.’s CSNet [74]. Since Dice loss is not suitable for small diffuse lesions, combining distribution-based loss with region-based loss has certainly helped.

# 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

# Conclusion and Future Works

While we included a fair number of studies in this review, the identified generalizability 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.

Further constructive and well-reported research is needed in this field, and it must be conducted on more data, perhaps by leveraging consortia (e.g., Human Connectome Project[³](#Footnote3), ENIGMA[4](#Footnote4)), and better data (e.g., generating structured labels from radiologist reports [102]).

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 [37]. For instance, deconvolution networks and guided back-propagation can explain the inner workings of DL networks [103,104].

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 [35].

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. They have started applying quantum algorithmic principles (e.g., running quantum operations on qubits) to ML [36], and to build expertise for when quantum hardware will be commercially available. This may increase computing speed significantly.

# Acknowledgments

**Author Contributions:** MB and MCVH conceived and presented the topic of this systematic review. MB and MCVH 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.

**Conflict of Interest Statement**: 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.

**Funding**: This systematic review and pilot analysis were funded by the University of Edinburgh.

# 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; **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; **y.o.**, Years Old.

# Appendix A. Supplementary Materials

Separate document composed of 5 sections:

1. Full search strategy
2. All visualizations / plots
3. Full data extraction & Study QA tables
4. All custom-built diagrams
5. General tips for developers

# Footnotes

1 <https://github.com/Elpazzu/UoE-Pilot-Analysis/>

² <https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools>

³ <https://www.humanconnectome.org/>

4 <https://enigma.ini.usc.edu/>

5 <https://www.prisma-statement.org/>

# Tables

## Table 1 – Study Selection Criteria

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Included** | **Excluded** | **Rationale** |
| Stroke types | Ischemic | Haemorrhagic | Differences in clinical presentations & etiologies |
| 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 offers better soft tissue contrast & resolution |
| Algorithms | * All learning approaches (e.g., supervised, unsupervised) * Algorithms segmenting **BOTH** ischemic 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 much better at learning complex hierarchical features * Synthetic data may not fully capture variations and complexities of real clinical stroke lesions |
| Population | * Humans (all ages/sexes) * All vascular risk factors | Non-human studies (e.g., animal-based) | Human-based studies are more clinically relevant and adhere to ethical guidelines regarding patient consent |
| Publishing | * Peer-reviewed studies * Proceedings of MICCAI, MIDL, and IEEE-led conferences **ONLY** * Publications in English * Publications between 2015-2023 | * Pre-prints * Studies **NOT** available in any of: MEDLINE, PubMed, IEEE Xplore, Web of Science, ScienceDirect and Springer | To only retain the most reliable sources of information and guarantee a certain level of quality |
| Completeness | Studies with sufficient information to be reproduced | Studies not reporting segmentation performance scores | Reproducibility is key in scientific research |

Table showing the inclusion and exclusion criteria related to the systematic review, along with the reasoning behind each of these criteria under “Rationale”. These criteria were used during the study selection process.

## Table 2 – Whole Group Analysis Methodology

|  |  |  |
| --- | --- | --- |
| **Effects models** | **Fixed-effects** | **Random-effects** |
| **Effect size** | Provided mean Dice scores, reported with their 95% confidence intervals (CI). | |
| **Effects estimates** | Weighted average of mean Dice scores. | |
| **Effects method** | **Inverse variance**: weights proportional to the inverse variance of study's Dice score. | **Heuristic method**: Tau² as ratio of Q-stat from fixed-effects to residual degrees of freedom. |
| **Null hypothesis (H0)** | **Homogeneity**: No significant difference in average mean Dice scores among studies. Any variation is due to random sampling error alone. | **Heterogeneity**: Significant difference in average mean Dice scores among studies, due not only to random sampling error, but also to variation in true effect sizes. |
| **Measurements** | **Heterogeneity**: Q-statistic (P-value), I²  **Overall effect test**: Z-score (P-value) | **Heterogeneity:** Q-statistic (P-value), I²  **Overall effect test**: Z-score (P-value) |

Table summarizing the statistical methods employed as part of the whole group analysis. Both fixed-effects and random-effects analyses were conducted.

## Table 3 – Subgroup Analysis Methodology

|  |  |
| --- | --- |
| **Rationale** | Make a more granular analysis of effect sizes. The 2 subgroups are:   * Studies using attention mechanisms * Studies not using attention mechanisms |
| **Statistical method** | The same approach as the one used for whole group analysis, this time using the random effects model only, and producing 1 set of measurements per subgroup.  A subgroup difference test was also performed using the “two-tailed” test. |

Table summarizing the statistical methods employed as part of the subgroup analysis, along with the rationale for conducting it. In this case, only a random-effects analysis was conducted.

## Table 4 – Sensitivity Analysis Methodology

|  |  |
| --- | --- |
| **Rationale** | Demonstrate the robustness of the conclusions produced from the meta-analysis by choosing Mean Precision scores as effect size, as this was the second most reported metric across retained studies. |
| **Effect size** | Provided mean **Precision** scores, reported with their 95% CI. |
| **Statistical method** | Precisely the same approach as the one used for the whole group analysis, but with a different effect size. |

Table summarizing the statistical methods employed as part of the sensitivity analysis, along with the rationale for conducting it. Both fixed-effects and random-effects analyses was conducted, and “Precision” was used as effect size instead of “Dice”.

## Table 5 – Meta-regression Analysis Methodology

|  |  |
| --- | --- |
| **Rationale** | Further assess whether there is statistically significant relationship between presence of attention mechanisms and likelihood of high mean Dice scores across studies. |
| **Statistical method** | **Model**: Ordinary Least Squares (OLS) (i.e., linear regression).  **Method**: Least squares. |
| **Variables** | **Independent**: Presence of attention mechanisms.  **Dependent**: Reported mean Dice scores. |

Table summarizing the statistical methods employed as part of the meta-regression analysis, along with the rationale for conducting it. The dependent variable used was the Dice score, and the independent variable was the presence of attention mechanisms.

## Table 6 – Publication Bias Analysis Methodology

|  |  |
| --- | --- |
| **Rationale** | Asses the presence of selective reporting, which may impact the results of the meta-analysis. |
| **Statistical methods** | **Funnel plot**:  x-axis: Mean Dice scores  y-axis: Standard Errors (SE)  **Egger’s test:**  Model: Robust Linear Model (RLM) (less sensitive to outliers than OLS)  Method: Reweighted least squares (IRLS) |

Table summarizing the statistical methods employed as part of the publication bias analysis, along with the rationale for conducting it. A funnel plot will be created, and Egger’s test will be conducted to verify the findings from the funnel plot.

## Table 7 – Sample Characteristics Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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. (1) | 64 | **MC**: 3 | Subacute | [18+] | - | - | - | 17.59 | [1.0, 346.1] |
| Gómez, S. | 75 | **MC**: 2 | Acute | [18+] | - | - | - | 37.83 | [1.6, 160.4] |
| Olivier, A. | 929 | **MC**: 6 | Acute Subacute | [16-94] | M: 63.7% F: 36.3% | 7.6 | 68.8h | 21.84 | - |
| Clèrigues, A. | 114 | **MC**: 4 | Acute Subacute | [18+] | - | - | - | SISS: 17.59  SPES: 133.21 | SISS: [1.0, 346.1]  SPES: [45.6, 252.2] |
| Liu, L. (1) | 64 | **MC**: 3 | Subacute | [18+] | - | - | - | 17.59 | [1.0, 346.1] |
| Moon, H. | 79 | - | Acute | - | M: 44.3% F: 55.7% | 9.3 | 83.8h | - | [0.0, 250] |
| Zhang, R. | 242 | **SC**: 1 | Acute | [35-90] | M: 60.3% F: 39.7% | - | - | - | - |
| Wong, K. | 875 | **SC**: 1 | Acute | - | M: 48.9% F: 51.1% | 6 | - | - | - |
| Khezrpour, S. | 64 | **MC**: 3 | Subacute | [18+] | - | - | - | 17.59 | [1.0, 346.1] |
| Hu, X. | 75 | **MC**: 2 | Acute | [18+] | - | - | - | 37.83 | [1.6, 160.4] |
| Gheibi, Y. | 44 | **MC**: 2 | Acute | - | - | - | - | - | *-* |
| Kumar, A. | 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. (2) | 79 | **MC**: 2 | Acute | [18+] | - | - | - | SPES: 133.21  LHC: - | SPES: [45.6, 252.2]  LHC: - |
| Zhao, B. | 582 | **SC**: 1 | Acute | - | - | - | - | - | - |
| Liu, C. | 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. (2) | 64 | **MC**: 3 | Subacute | [18+] | *-* | *-* | *-* | 17.59 | [1.0, 346.1] |
| Liu, L. (3) | 114 | **MC**: 4 | Acute Subacute | [18+] | - | - | - | SISS: 17.59  SPES: 133.21 | SISS: [1.0, 346.1]  SPES: [45.6, 252.2] |
| Aboudi, F. | 64 | **MC**: 3 | Subacute | [18+] | *-* | *-* | *-* | 17.59 | [1.0, 346.1] |
| Pinto, A. | 75 | **MC**: 2 | Acute | [18+] | - | - | - | 37.83 | [1.6, 160.4] |
| Choi, Y. | 54 | **MC**: 2 | Acute | [18+] | - | - | - | 37.83 | [1.6, 160.4] |
| Kim, Y. | 296 | **SC**: 1 | Acute | [58–79] | M: 61.3% F: 38.7% | 2.3 | 12.7h | 12.19 | [0.0, 279.4] |
| Woo, I. | 429 | **SC**: 1 | Acute | [24-98] | M: 62.3% F: 37.7% | - | 21.4h | - | - |
| Lee, A. | 429 | **SC**: 1 | Acute | [24-98] | M: 62.3% F: 37.7% | - | 21.4h | 27.44 | [0.3, 227.6] |
| Lee, S. | 472 | **SC**: 1 | Acute | [19+] | M: 63.3% F: 36.7% | 3 | 4.9h | 3.62 | [0.52, 71.8] |
| Karthik, R. (3) | 64 | **MC**: 3 | Subacute | [18+] | *-* | *-* | - | 17.59 | [1.0, 346.1] |
| Zhang, L. | 64 | **MC**: 3 | Subacute | [18+] | *-* | *-* | - | 17.59 | [1.0, 346.1] |
| Ou, Y. | 99 | **SC**: 1 | Acute | - | - | - | - | - | - |
| Vupputuri, A. | 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 | **MC**: 3 | Subacute | [18+] | *-* | *-* | *-* | 17.59 | [1.0, 346.1] |
| Duan, W. | 120 | **SC**: 1 | Acute | - | - | - | - | - | - |
| Lucas, C. | 75 | **MC**: 2 | Acute | [18+] | - | - | - | 37.83 | [1.6, 160.4] |
| Nazari-Farsani, S. | 445 | **MC**: 6+ | Acute | - | M: 50% F: 50% | 13 | 6.2h | 50 | [15, 123] |
| Wei, Y. | 216 | **SC**: 1 | Acute | - | M: 69.7% F: 30.3% | - | - | - | - |
| Li, C. | 60 | **SC**: 1 | Acute | [49-88] | - | - | - | - | - |
| Liu, Z. | 212 | **SC**: 1 | Acute Subacute | - | M: 62% F: 38% | - | - | - | - |
| Cornelio, L. | 75 | **MC**: 2 | Acute | [18+] | - | - | - | 37.83 | [1.6, 160.4] |
| Yu, Y. | 182 | **MC**: 6+ | Acute | - | M: 46.7% F: 53.3% | 15 | - | 54 | [16, 117] |
| Wu, Z. | 400 | **MC**: 3 | Subacute | [18+] | - | - | - | 27.94 | [0.0575, 340.28] |
| Guerrero, R. | 250 | **SC**: 1 | Acute | - | - | - | - | - | - |

Table summarizing the sample characteristics data extracted from retained papers. The second row is used to provide further explanations or precisions when more clarity is needed.

## Table 8 – Medical Imaging Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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. (1) | 2-H | **MM**: {FLAIR, T2WI, T1WI, DWI-b1000} | 2D | DWI-FLAIR | 3T |
| Gómez, S. | 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. | Not reported | **SM**: {DWI-b0, DWI-b1000, DWI-ADC} | 3D | None reported | 1.5T 3T |
| Clèrigues, A. | 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. (1) | 2-H | **MM**: {FLAIR, DWI-b1000} | 2D | DWI-FLAIR | 3T |
| Moon, H. | 1-VH | **MM**: {FLAIR, DWI-b1000} | 2D | None reported | 1.5T |
| Zhang, R. | 3-M | **SM**: {DWI-b0, DWI-b1000, DWI-ADC} | 3D | None reported | 1.5T 3T |
| Wong, K. | Not reported | **SM**: {DWI-b0, DWI-b1000, DWI-eADC} | 2D | None reported | 1.5T 3T |
| Khezrpour, S. | 2-H | **SM**: {FLAIR} | 2D | DWI-FLAIR | 3T |
| Hu, X. | 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. | Not reported | **MM**: {FLAIR, DWI} | 2D | None reported | - |
| Kumar, A. | 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. (2) | 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. | 2-H | **SM**: {DWI-ADC, DWI-b0, DWI-b1000} | 2D | None reported | 1.5T 3T |
| Liu, C. | 3-M | **SM**: {DWI-b0, DWI-ADC, DWI-IS} | 3D | None reported | 1.5T 3T |
| Karthik, R. (2) | 2-H | **MM**: {FLAIR, T2WI, T1WI, DWI-b1000} | 2D | DWI-FLAIR | 3T |
| Liu, L. (3) | 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. | 2-H | **MM**: {FLAIR, T2WI, T1WI, DWI-b1000} | 2D | DWI-FLAIR | 3T |
| Pinto, A. | 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. | 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. | 1-VH | **SM**: {DWI-b0, DWI-b1000, DWI-ADC} | 2D | None reported | 1.5T 3T |
| Woo, I. | 1-VH | **SM**: {DWI-b1000, DWI-b0, DWI-ADC} | 2D | None reported | 1.5T 3T |
| Lee, A. | 1-VH | **SM**: {DWI-b1000, DWI-b0, DWI-ADC} | 2D | None reported | 1.5T 3T |
| Lee, S. | 3-M | **MM**: {DWI, DWI-ADC, FLAIR, PWI-Tmax, PWI-TTP, Pred(init)} | 3D | DWI-PWI | 1.5T 3T |
| Karthik, R. (3) | 2-H | **MM**: {FLAIR, T2WI, T1WI, DWI-b1000} | 2D | DWI-FLAIR | 3T |
| Zhang, L. | 2-H | **SM**: {DWI-b1000} | 2D | DWI-FLAIR | 3T |
| Ou, Y. | 1-VH | **SM**: {DWI-b1000, DWI-eADC} | 2.5D | None reported | 1.5T 3T |
| Vupputuri, A. | 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. | 2-H | **MM**: {FLAIR, T2WI, T1WI, DWI-b1000} | 2D | DWI-FLAIR | 3T |
| Duan, W. | Not reported | **MM**: {T2WI, DWI-b1000, DWI-b0} | 3D | None reported | - |
| Lucas, C. | 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. | Not reported | **SM**: {DWI-b1000, DWI-ADC} | 3D | None reported | 1.5T 3T |
| Wei, Y. | 1-VH | **SM**: {DWI-b1000} | 2D | T1-T2;  T2-FLAIR;  T1-FLAIR | 3T |
| Li, C. | 1-VH | **MM**: {T1WI, T2WI, T2WI-FLAIR, DWI, DWI-ADC} | 2D | T2-FLAIR;  T1-FLAIR | 1.5T |
| Liu, Z. | Not reported | **MM**: {T2WI, DWI, DWI-ADC} | 2D | None reported | 1.5T 3T |
| Cornelio, L. | 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. | Not reported | **MM**: {DWI-b1000, DWI-ADC, PWI-Tmax, PWI-MTT, PWI-CBF, PWI-CBV} | 2.5D | None reported | 1.5T 3T |
| Wu, Z. | 1-VH | **MM**: {DWI-b1000, DWI-ADC, FLAIR} | 2D | DWI-FLAIR | 1.5T 3T |
| Guerrero, R. | 1-VH | **MM**: {FLAIR, T1WI} | 2D | None reported | 1.5T |

Table summarizing the medical imaging data extracted from retained papers. The second row is used to provide further explanations or precisions when more clarity is needed.

## Table 9 – Data Pre-processing Techniques Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **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. (1) | ISLES2015 SISS | - Normalization  - Skull-stripping | - | - Resizing | - Registration | - |
| Gómez, S. | ISLES2017 | - Normalization  - Contrast adjustment  - Skull-stripping | - | - Resizing  - Rescaling | - Registration | - |
| Olivier, A. | Proprietary | - Normalization | - | - Rescaling - Zero-padding - Cropping | - | - |
| Clèrigues, A. | ISLES2015 SISS ISLES2015 SPES | - Normalization - Skull-stripping | - | - | - Registration | - |
| Liu, L. (1) | ISLES2015 SISS | - Skull-stripping | - | - | - Registration | - |
| Moon, H. | Proprietary | - Normalization | - Skull-stripping | - Zero-padding - Resizing | - Registration | - |
| Zhang, R. | Proprietary | - Normalization | - | - Zero-padding - Cropping - Resizing | - | - |
| Wong, K. | Proprietary | - Normalization | - | - | - | - |
| Khezrpour, S. | ISLES2015 SISS | - Contrast adjustment - RGB to greyscale - Skull-stripping | - | - Cropping - Resizing | - Registration | - |
| Hu, X. | ISLES2017 | - Skull-stripping | - | - Resizing - Cropping | - Registration | - |
| Gheibi, Y. | Proprietary | - | - | - Zero-padding | - Splitting into 2D | - |
| Kumar, A. | ISLES2015 SPES ISLES2015 SSIS  ISLES2017 | - Normalization - Skull-stripping |  | - Resizing | - Converting to 3D - Registration | - Slice classification |
| Liu, L. (2) | ISLES2015 SPES  Proprietary | - Normalization - Smoothing - Skull-stripping | - | - | - Registration | - |
| Zhao, B. | Proprietary | - Normalization | - | - | - | - |
| Liu, C. | Proprietary | - Normalization | - | - Resizing - Rescaling | - | - Slice classification - Skull-stripping |
| Karthik, R. (2) | ISLES2015 SISS | - Skull-stripping | - | - | - Registration | - |
| Liu, L. (3) | ISLES2015 SPES ISLES2015 SISS | - Normalization - Skull-stripping | - | - Zero-padding - Cropping - Resizing | - Registration - Splitting into 2D | - |
| Aboudi, F. | ISLES2015 SISS | - RGB to greyscale - Skull-stripping | - | - Resizing - Rescaling | - Registration | - |
| Pinto, A. | ISLES2017 | - Normalization - Bias field correction - Skull-stripping | - | - Resizing | - Registration | - |
| Choi, Y. | ISLES2016 | - Normalization - Skull-stripping | - | - Resizing - Rescaling | - Registration | - |
| Kim, Y. | Proprietary | - Normalization | - | - Resizing | - | - |
| Woo, I. | Proprietary | - Normalization | - | - | - Registration | - |
| Lee, A. | Proprietary | - Normalization | - | - Resizing | - Registration | - |
| Lee, S. | Proprietary | - | - | - Resizing - Rescaling | - Registration | - |
| Karthik, R. (3) | ISLES2015 SISS | - Normalization - Skull-stripping | - | - Cropping - Rescaling | - Registration | - Slice classification |
| Zhang, L. | ISLES2015 SISS | - Normalization - Skull-stripping | - | - Cropping - Rescaling | - Registration | - Slice classification |
| Ou, Y. | Proprietary | - Normalization - Skull-stripping | - | - Resizing | - | - |
| Vupputuri, A. | ISLES2015 SPES ISLES2015 SISS  ISLES2017 (IS17) | - RGB to greyscale - Normalization - Skull-stripping | - | - | - Registration | - |
| Abdmouleh, N. | ISLES2015 SISS | - Normalization - Skull-stripping | - | - | - Registration | - |
| Duan, W. | Proprietary | - Normalization - Skull-stripping | - | - Resizing | - | - |
| Lucas, C. | ISLES2017 | - Skull-stripping | - | - Rescaling | - Registration | - |
| Nazari-Farsani, S. | UCLA  iCAS  DEFUSE  DEFUSE-2 | - Normalization | - | - | - Registration | - |
| Wei, Y. | Proprietary | - Skull-stripping | - | - Rescaling | - Registration | - |
| Li, C. | Proprietary | - | - | - Resizing | - | - |
| Liu, Z. | Proprietary | - Normalization - Skull-stripping | - | - Cropping - Resizing | - Registration | - |
| Cornelio, L. | ISLES2017 | - RGB to greyscale - Contrast adjustment - Normalization - Skull-stripping | - | - Resizing | - Registration | - |
| Yu, Y. | iCAS  DEFUSE-2 | - Normalization | - | - | - Registration | - |
| Wu, Z. | ISLES2022 | - Skull-stripping | - | - Resizing - Rescaling | - Registration | - |
| Guerrero, R. | Proprietary | - Normalization | - | - Resizing | - Registration | - |

Table summarizing the data pre-processing techniques extracted from retained papers. The third row is used to provide further explanations or precisions when more clarity is needed.

## Table 10 – Deep Learning Architectures Data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **First author** | **Architecture** | **Loss function** | **Segmentation type** | **Activation functions** | **Regularization method** | **Optimization method** | **Epochs** |
| Karthik, R. (1) | U-Net | Dice | Semantic | Leaky ReLU, ReLU, Softmax | Data augmentation | Adam | 120 |
| Gómez, S. | U-Net | Focal | Semantic | ReLU, Sigmoid | - Data augmentation  - Weight decay  - Class weighting | AdamW | 600 |
| Olivier, A. | U-Net | Dice | Semantic | Leaky ReLU, Softmax | - Data augmentation  - ES on validation loss | Adam | - |
| Clèrigues, A. | U-Net | Focal | Semantic | PReLU, Softmax | - Data augmentation  - ES on MAE/L1 loss  - Dropout  - Class weighting | AdaDelta | - |
| Liu, L. (1) | U-Net | Dice | Semantic | ReLU, Sigmoid | - Data augmentation  - Dropout | Adam | 150 |
| Moon, H. | U-Net | BCE | Semantic | ReLU, Sigmoid | - | Adam | 200 |
| Zhang, R. | DenseNet | Dice | Semantic | ReLU, Softmax | - Data augmentation  - Weight decay  - Learning rate adjust. | SGD | 2000 |
| Wong, K. | U-Net | Dice | Semantic | ReLU, ? | Data augmentation | - | - |
| Khezrpour, S. | U-Net | Dice | Semantic | ReLU, Sigmoid | - Data augmentation  - ES on validation loss | Adam | - |
| Hu, X. | U-Net + ResNet | Focal | Semantic | ReLU, Sigmoid | - Data augmentation  - Class weighting | Adam | 1500 |
| Gheibi, Y. | U-Net + ResNet | Custom | Semantic | ReLU, Sigmoid | - Data augmentation  - Weight decay  - Dilution | Adam | - |
| Kumar, A. | U-Net | BCE-Dice | Semantic | ReLU, Softmax | - Data augmentation  - Dropout  - ES on validation set  - Learning rate adjust. | Adam | 200 |
| Liu, L. (2) | U-Net + ResNet | Custom | Semantic | Leaky ReLU, Sigmoid | Data augmentation | - | 70 |
| Zhao, B. | CNN | BCE | Semantic | ReLU, Sigmoid | - Data augmentation  - ES on validation loss | RAdam | - |
| Liu, C. | U-Net | BCE-Dice | Semantic | SeLU (Self-normalized), Sigmoid | - Weight decay  - ES on training & val.  - Learning rate adjust.  - Class weighting | Adam | 200 |
| Karthik, R. (2) | U-Net | Dice | Semantic | ReLU, Sigmoid | Data augmentation | - | 150 |
| Liu, L. (3) | U-Net + DenseNet | CE-Dice | Semantic | ReLU, Sigmoid | - Data augmentation  - Dropout | Adam | 8 |
| Aboudi, F. | U-Net | CE | Semantic | ReLU, Sigmoid | Data augmentation | Adam | 100 |
| Pinto, A. | U-Net | Dice | Semantic | - | - | Adam | - |
| Choi, Y. | U-Net + CNN + ResNet | CE-Dice | Semantic | ReLU, Softmax | - Data augmentation  - Weight decay  - Dropout  - *ES on ?* | Adam | - |
| Kim, Y. | U-Net | Dice | Semantic | ReLU, Sigmoid | - | Adam | 1000 |
| Woo, I. | U-Net + DenseNet | - | Semantic | ReLU, Sigmoid | - | - | *-* |
| Lee, A. | U-Net | Dice | Semantic | ReLU, Sigmoid | - | - | *-* |
| Lee, S. | U-Net | Dice | Semantic | ReLU, Sigmoid | ES on validation loss | Adam | - |
| Karthik, R. (3) | U-Net | Dice-CE + Softmax-CE | Semantic | ReLU, Softmax | - Data augmentation  - Masked dropout  - Dropout  - Learning rate adjust. | Adam | 150 |
| Zhang, L. | U-Net | CE | Semantic + Instance | ReLU, Softmax | - Data augmentation  - Momentum  - Weight decay | SGD | - |
| Ou, Y. | U-Net | BCE | Semantic | ReLU, Softmax | - | RMSprop | 100 |
| Vupputuri, A. | U-Net | BCE | Semantic | Leaky ReLU, Softmax | - ES on validation set  - Dropout | Adam | 30 |
| Abdmouleh, N. | U-Net | CE | Semantic | ReLU, Sigmoid | Data augmentation | Adam | 20 |
| Duan, W. | CNN + ResNet | Dice-CE | Semantic | PReLU, Softmax | Data augmentation | Adam | 600 |
| Lucas, C. | U-Net | Soft QDice | Semantic | ReLU, Sigmoid | Data augmentation | Adam | 100 |
| Nazari-Farsani, S. | U-Net | BCE-Volume-MAE-Dice | Semantic | ReLU, Sigmoid | - Data augmentation  - Class weighting  - Dropout | Adam | 80 |
| Wei, Y. | U-Net + ResNet | Focal Tversky | Semantic | ReLU, Softmax | - Data augmentation  - Class weighting  - Learning rate adjust. | Adam | 150 |
| Li, C. | U-Net | CE | Instance | ReLU, Sigmoid | - Data augmentation  - Class weighting | SGD | 200 |
| Liu, Z. | CNN + ResNet | Dice | Semantic | ReLU, Sigmoid | - Data augmentation  - Weight decay  - Learning rate adjust. | Adam | 500 |
| Cornelio, L. | U-Net | Dice | Semantic | ReLU, Sigmoid | - Dropout  - Weight decay | Adam | 50 |
| Yu, Y. | U-Net | BCE-Volume-MAE-Dice | Semantic | ReLU, Sigmoid | - Data augmentation  - Class weighting  - Dropout | Adam | 120 |
| Wu, Z. | U-Net + MLP | Dice + Boundary | Semantic | ReLU, Softmax | - Weight decay  - Learning rate adjust. | AdamW | 35 |
| Guerrero, R. | U-Net + ResNet | CCE | Semantic | ReLU, Softmax | - Data augmentation  - Class weighting  - Learning rate adjust. | Adam | - |

Table summarizing the deep learning architectures data extracted from retained papers.

## Table 11 – Attention Mechanisms Data

|  |  |  |
| --- | --- | --- |
| **First author** | **Attention mechanism** | **Attention Type** |
| Karthik, R. (1) | None | N/A |
| Gómez, S. | Additive cross-attention | Spatial attention |
| Olivier, A. | None | N/A |
| Clèrigues, A. | None | N/A |
| Liu, L. (1) | Self-gated soft attention | Hybrid attention |
| Moon, H. | None | N/A |
| Zhang, R. | None | N/A |
| Wong, K. | None | N/A |
| Khezrpour, S. | None | N/A |
| Hu, X. | None | N/A |
| Gheibi, Y. | None | N/A |
| Kumar, A. | None | N/A |
| Liu, L. (2) | None | N/A |
| Zhao, B. | Squeeze- excitation | Channel attention |
| Liu, C. | Dual attention gates | Hybrid attention |
| Karthik, R. (2) | Attention gates | Spatial attention |
| Liu, L. (3) | None | N/A |
| Aboudi, F. | None | N/A |
| Pinto, A. | None | N/A |
| Choi, Y. | None | N/A |
| Kim, Y. | None | N/A |
| Woo, I. | Squeeze- excitation | Channel attention |
| Lee, A. | Squeeze-excitation | Channel attention |
| Lee, S. | None | N/A |
| Karthik, R. (3) | Multi-residual attention | Hybrid attention |
| Zhang, L. | None | N/A |
| Ou, Y. | None | N/A |
| Vupputuri, A. | Multi-path attention | Hybrid attention (includes self-attention) |
| Abdmouleh, N. | None | N/A |
| Duan, W. | None | N/A |
| Lucas, C. | None | N/A |
| Nazari-Farsani, S. | Attention gates | Spatial attention |
| Wei, Y. | None | N/A |
| Li, C. | None | N/A |
| Liu, Z. | None | N/A |
| Cornelio, L. | None | N/A |
| Yu, Y. | Attention gates | Spatial attention |
| Wu, Z. | Multi-head self-attention | Hybrid attention (includes self-attention) |
| Guerrero, R. | None | N/A |

Table summarizing the attention mechanisms data extracted from retained papers.

## Table 12 – Performance and Generalizability Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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. (1) | 0.701 | - | - | - | N | L | 7h30 | **CPU**: 3.6GHz QuadCore Intel Gen7  **RAM**: 32GB  **GPU**: Nvidia Quadro P4000  **Library**: Keras/TensorFlow |
| Gómez, S. | 0.36 ± 0.21 | 0.42 ± 0.25 | 0.48 ± 0.29 | - | N | L | - | - |
| Olivier, A. | 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. | 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. (1) | 0.764 | - | 0.944 | 3.19 | N | M | - | - |
| Moon, H. | 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. | 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. | 0.84 ± 0.03 | 0.84 ± 0.03 | 0.89 ± 0.03 | - | N | M | - | - |
| Khezrpour, S. | 0.852 | 0.998 | 0.856 | - | N | L | - | **GPU**: Google Cloud Compute (K80)  **Library**: Keras/TensorFlow |
| Hu, X. | 0.30 ± 0.22 | 0.35 ± 0.27 | 0.43 ± 0.27 | - | N | L | - | **GPU**: 4 x Nvidia Titan Xp |
| Gheibi, Y. | 0.792 | - | - | - | N | M | 1h27 | **GPU**: Nvidia Tesla P100  **Library**: Keras |
| Kumar, A. | - | 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. (2) | 0.817 | - | - | 1.92 | N | M | 0h36 | **Library**: Keras/TensorFlow |
| Zhao, B. | 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. | 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. (2) | 0.7535 | - | - | - | N | L | 34h04 | **CPU**: 3.6GHz QuadCore Intel (Gen 7)  **RAM**: 32GB  **GPU**: Nvidia Quadro P4000  **Library**: Keras/TensorFlow |
| Liu, L. (3) | 0.68 ± 0.19 | - | - | 39.975 ± 27.95 | N | M | - | - |
| Aboudi, F. | 0.558 | 0.998 | - | - | N | L | - | **CPU**: Intel Core i5 8th gen  **RAM**: 8GB  **GPU**: Nvidia GeForce GTX 1050  **Library**: Keras/TensorFlow |
| Pinto, A. | 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. | 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. | 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. | 0.858 ± 0.0734 | - | - | - | Dice: - S (<10mL): 0.82 - L (>10mL): 0.89 | L | - | - |
| Lee, A. | 0.854 ± 0.008 | 0.845 | 0.995 | - | N | L | - | - |
| Lee, S. | 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. (3) | 0.775 | 0.751 | 0.801 | - | N | L | - | **CPU**: 4 cores  **OS**: Ubuntu 16.04  **RAM**: 32GB  **GPU**: 2 x Nvidia Tesla P100  **Library**: PyTorch |
| Zhang, L. | 0.433 | - | 0.356 | - | N | L | - | **GPU**: Nvidia GeForce GTX 1080 Ti  **Library**: Keras/TensorFlow |
| Ou, Y. | 0.865 | 0.894 | 0.818 | - | N | M | 4h | **GPU**: 4 x Nvidia Quadro RTX 6000  **Library**: PyTorch |
| Vupputuri, A. | 0.71 | - | 0.897 | - | N | M | - | **GPU**: Nvidia Tesla K80 |
| Abdmouleh, N. | 0.71 ± 0.11 | - | - | - | N | L | - | - |
| Duan, W. | 0.677 ± 0.165 | - | - | 85.462 ± 14.496 | N | M | - | **GPU**: Nvidia GTX 1080 Ti  **Library**: PyTorch |
| Lucas, C. | 0.35 | 0.52 | 0.35 | 21.48 | N | L | - | **GPU**: Nvidia Titan Xp (12GB)  **Library**: PyTorch |
| NazariFarsani S. | 0.5 | - | 0.6 | - | N | M | - | - |
| Wei, Y. | 0.828 | - | - | - | Dice:  S (<769 pixels): 0.761 L (>769): 0.83 | M | - | - |
| Li, C. | - | - | - | 38.27mm | N | L | - | - |
| Liu, Z. | 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. | 0.34 | - | - | - | N | L | 5h | **OS**: Ubuntu v.16.04.3  **GPU**: Nvidia GeForce GTX  **Library**: Keras/TensorFlow |
| Yu, Y. | 0.53 | 0.53 | 0.66 | - | N | M | 35h | **GPU**: Nvidia Quadro GV100 & Nvidia Tesla V100-PCIE  **Library**: Keras/TensorFlow |
| Wu, Z. | 0.856 | 0.883 | 0.854 | 27.34 | N | M | 0h21 | **GPU**: 6 x Nvidia Tesla 4s  **Library**: PyTorch |
| Guerrero, R. | 0.4±0.252 | - | - | - | N | L | - | **Library**: Lasagne/Theano |

Table summarizing the performance and generalizability data from retained papers. The second row is used to provide further explanations or precisions when more clarity is needed.

## Table 13 – Pilot Analysis Results Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 |

Table summarizing the performance (i.e., Mean Dice scores) of the deployed algorithms as part of the pilot analysis. Performance are reported on both the training and validation sets. 24 experiments were conducted in total, by mixing up:

* 4 architectures: UResNet50, UNet, AG-UResNet50, AG-UNet, where AG means attention-gated
* 3 loss functions: Dice loss, Compound 0.5Dice+0.5BCE loss, Compound 0.7Dice+0.3BCE loss
* 2 input image sequences: Single-modality approach (DWI), Multi-modality approach (DWI, FLAIR, T1-WI, and T2-WI)

## Table 14 – Summary of Average Training Times

|  |  |
| --- | --- |
| **Algorithm** | **Average training time** |
| UResNet50 | 5h43 (SM) ; 8h01 (MM) |
| UNet | 5h31 (SM); 8h26 (MM) |
| AG-UResNet50 | 6h15 (SM); 9h10 (MM) |
| AG-UNet | 5h55 (SM); 8h41 (MM) |

Table summarizing, per model architecture, the average times when training the models across all the experiments. “SM” refers to single-modality experiments, and “MM” to multi-modality experiments.

# Figures

## Figure 1 – Existing Similar Reviews Heatmap

A screenshot of a document

Description automatically generated

Heatmap showing the differences in scope between this present systematic review (i.e., Baaklini2024) and existing reviews covering similar topics. The legend of the heatmap and other useful information are displayed at the bottom of the image.

While several related reviews have been published since 2017 [1-12,16,73,82,83], their scopes differ from ours. Our review is unique in the way it supplemented the meta-analysis by a pilot analysis as well as an in-depth meta-analysis which included whole group, subgroup, sensitivity and meta-regression analyses.

## Figure 2 – Architecture of Main Types of Blocks

A diagram of a block diagram

Description automatically generated

Image illustrating the internal architecture of standard blocks that may compose deep learning networks. The colour coding is needed to better understand the contents of this image.:

* **a) Architecture of a standard convolution block** (typically included in CNNs): a standard convolution block in a CNN consists of a convolution operation on a 3x3 kernel, which produces a feature map on which batch normalization is applied, before passing through a ReLU activation function to introduce non-linearity
* **b) Architecture of a standard residual block** (typically includes in ResNets): a standard ResNet block consists of an “identity path” (green arrow) that can bypass the “residual path”, thus giving the network the option to simply copy activations from layer to layer and preserve information when learned features do not require more depth. The residual path consists of 2 consecutive convolution blocks. Whether the identity of the residual path was taken, BN and ReLU activation are applied before passing to the next network block
* **c) Architecture of a standard dense block** (typically included in DenseNets): It employs a radical dense connection strategy, interconnecting all layers to maximize information and gradient propagation, as seen with the green, red and blue arrows. Concretely, it consists of 3 consecutive convolution blocks with a slightly different sequence of operations(i.e., BN, then ReLU, then the convolution operation)

For better readability, please download the file provided in [Appendix.A.D.](#AppendixA)

## Figure 3 – Architecture of Attention Mechanisms

A diagram of a flowchart

Description automatically generated

Image illustrating the internal architecture of common attention mechanism blocks that may be incorporated in deep learning networks. The colour coding is not essential to understand the contents of this image, it was only used to make it simpler for the reader:

* a) Architecture of a standard spatial attention block: the feature map first goes through a max-pooling and an average pooling operation, in parallel. The outputs of the 2 are then concatenated using element-wise addition. A 3x3 convolution then a Sigmoid activation are then applied, in order to produce a binary spatial attention map, where only supposedly important pixels are in the foreground, and supposedly unimportant ones in the background. The original feature map and the spatial attention map are then combined using element-wise multiplication, in order to produce the final attention-enhanced feature map that is passed to the next layer of the network
* b) Architecture of a standard channel attention block: the feature map first goes through a max-pooling and an average pooling operation, in parallel. Each output goes through 2 consecutive fully-connected layers, in order to produce an average-pooled weight map and a max-pooled weight map. The 2 maps are then concatenated using element-wise addition, before going through sigmoid activation in order to produce the channel attention weight map, which either classifies a feature map as unimportant for learning (Wo=0) or important for learning (Wo=1). The original feature map and the channel attention weight map are then concatenated using element-wise multiplication in order to produce the final attention-enhanced feature map that is passed to the next layer of the network

For better readability, please download the file provided in [Appendix.A.D.](#AppendixA)

## Figure 4 – Architecture of “AG-UResNet50”

A computer screen shot of a diagram

Description automatically generated

A computer screen shot of a computer

Description automatically generated

A computer screen shot of a computer

Description automatically generated

Image illustrating, in landscape mode, the architecture of “AG-UResNet50”, which was deployed as part of the pilot analysis. For better readability, please download the file provided in [Appendix.A.D.](#AppendixA)

AG-UResNet50 is a 5-level end-to-end U-Net, with as main novelty the replacement of its encoder path by ResNet50 (as-is). Using U-Net in combination with ResNet50 allows us to leverage the power of skip connections even further, and to make the network deeper.

Max-pooling was only 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 2 were used for remaining down-sampling operations on the encoder side, in order to better retain image details.

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. We also kept the up-sampling interpolation algorithm, which 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.

## Figure 5 – Architecture of Attention Gate in “AG-UResNet50”

A screenshot of a computer screen

Description automatically generated

Image illustrating the architecture of every attention gate used in “AG-UResNet50”.

An AG takes 2 input vectors, Wg and Wx, that are summed element-wise, resulting in aligned weights becoming larger and unaligned weights smaller. The output vector then goes through ReLU activation, 1x1 convolution, and sigmoid activation in order to scale the vector in the 0-1 range, and 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 then passed along in the skip connection.

## Figure 6 – Study Selection Flow Diagram

A screenshot of a flowchart

Description automatically generated

Image illustrating the study selection flow related to our systematic review. Among 1485 originally identified papers, only 39 were retained in our systematic review, and only 17 in the meta-analysis. Along the way, duplicated were removed, and articles were excluded following each of the title/abstract screening and the full-text screening steps. 1 article was found and added during a random search conducted after the full search.

## Figure 7 – Sample Characteristics Plots

A close-up of a graph

Description automatically generated

**(a) (b)**

A screenshot of a graph

Description automatically generated

**(c) (d)**

Image illustrating some of the plots we created as part of the meta-analysis, in order to analyse sample characteristics data extracted from retained papers:

* a) **Gender proportions** (reported for 13/39 studies): males were generally over-represented (Mean:58%) except in a few studies where the opposite was true (e.g., Moon et al. [65]).
* b) **Mean stroke-to-MRI times** (reported for 8/39 studies): reported patient mean TSS were exclusively in the acute interval, with 2 studies very close to the hyperacute-acute border
* c) **Sample sizes** (reported for 39/39 studies): only 4 papers used sample sizes above 500 (Mean:252.2)
* d) **NIHSS mean scores** (reported for 8/39 studies): reported mean NIHSS [45] were always on the “minor” or “moderate” ranges

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 8 – Medical Imaging Data Plots

A graph of a graph of a graph

Description automatically generated with low confidence

**(a) (b)**

Image illustrating the plots we created as part of the meta-analysis, in order to analyse medical imaging data extracted from retained papers:

* a) **Correlation of Image dimensions with Image spatial resolution**: 2D models exclusively used high or very high-resolution images, and 3D models exclusively used moderate or low-resolution images
* b) **Correlation of Image dimensions with Model training mode**: 3D models adopted patch-wise training in 8/10 studies

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 9 – Model Architecture Types

A white rectangular object with black text

Description automatically generated

Image illustrating the count of model architecture types used across retained studies. 35/39 studies deployed U-Net-based models. ResNet-based models were used in 8 studies, and DenseNets in 3 studies. U-Net was combined with other architecture types in 10 studies.

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 10 – Model Performance and Generalizability Plots

A graph of a graph and a graph of a graph

Description automatically generated with medium confidence

1. **(b)**

A diagram of a graph

Description automatically generated with medium confidence

**(c) (d)**

A close-up of a graph

Description automatically generated

. **(e) (f)**

Image illustrating some of the plots we created as part of the meta-analysis, in order to analyse model performance and generalizability data extracted from retained papers:

* a) **Correlation between Dice scores and Generalizability scores**: there is a positive correlation
* b) **Correlation between Dice scores and Precision scores**: there is a positive correlation
* c) **Correlation between Dice scores and Mean lesion volumes, with image spatial resolution as third dimension:** very high resolution images were used when mean lesion volumes were relatively small (< 35ml)
* d) **Correlation between Dice scores and Spatial resolution:** there is a positive correlation
* e) **Correlation between Dice scores and Total number of layers in the model architecture, with model types as third dimension:** when U-Nets were deeper, Dice scores were higher
* f) **Correlation between Dice scores and Total number of layers in the model architecture, with the presence/absence of attention as third dimension:** when attention-based networks were deeper, Dice scores were higher

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 11 – Whole Group Forest Plot

A screenshot of a graph

Description automatically generated

Image illustrating the forest plot we created as part of the whole group analysis in the meta-analysis. It shows per study, the Dice, STD, sample size, and weight assigned from both the fixed-effects (“F. Weight(%)”) and random-effects analyses (“R. Weight(%)”). 17 studies were represented:

* Diamonds represent point estimates. Horizontal lines going through them represent 95% CIs
* The line of null effect was not plotted, as in this case, there is no “expected” Dice score under the null hypothesis
* Studies are ordered from highest to lowest mean Dice scores

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 12 – Subgroups Forest Plot

A screenshot of a graph

Description automatically generated

Image illustrating the forest plot we created as part of the subgroup analysis in the meta-analysis. It shows per study, the Dice, STD, sample size, and weight assigned from both the fixed-effects (“F. Weight(%)”) and random-effects analyses (“R. Weight(%)”). It also separated studies which used attention (“Y”), which are 5 in total from those not using attention (“N”), which are 12 in total:

* Diamonds represent point estimates. Horizontal lines going through them represent 95% CIs
* The line of null effect was not plotted, as in this case, there is no “expected” Dice score under the null hypothesis
* Within each subgroup, studies are ordered from highest to lowest mean Dice scores

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 13 – Sensitivity Analysis Forest Plot

A screenshot of a computer screen

Description automatically generated

Image illustrating the forest plot we created as part of the sensitivity analysis in the meta-analysis. It shows per study, the **Precision** **score**, STD, sample size, and weight assigned from both the fixed-effects (“F. Weight(%)”) and random-effects analyses (“R. Weight(%)”). 8 studies are represented:

* Diamonds represent point estimates. Horizontal lines going through them represent 95% CIs
* The line of null effect was not plotted, as in this case, there is no “expected” Precision score under the null hypothesis
* Studies are ordered from highest to lowest mean Precision scores

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 14 – Meta-regression Summary Statistics

A screenshot of a computer

Description automatically generated

Image showing the summary statistics related to the meta-regression conducted as part of the meta-analysis. It assesses the statistical significance of the relationship between “attention mechanisms” and “Dice scores”:

* 8.1% of the variance in Dice scores is explained by the presence of attention (R-squared: 0.081)
* The slope indicating the change in Dice associated with the presence of attention is not statistically significant (0.117,p=0.27>0.05), therefore we cannot conclude that the presence of attention has a significant impact on the likelihood of high Dice
* The 95% CI of the slope [-0.100,0.334] includes 0, suggesting that the effect of attention on Dice is not statistically significant at p=0.05

## Figure 15 – Funnel Plot

A graph with red and blue dots

Description automatically generated

Image illustrating the funnel plot we created as part of the publication analysis in the meta-analysis:

* Blue points represent individual studies
* Red diagonals represent lower and upper bounds of 95% CIs
* Green vertical line represents the weighted mean of reported mean Dice scores
* The y-axis shows the Standard Errors (SE), and x-axis shows the mean Dice scores

The funnel plot is asymmetric, suggesting **presence of publication bias**. The high concentration of data points on the right side of the green line suggests selective reporting of studies with high Dice. Specifically, the 6 points with very low SE (upper-right corner) represent the studies that were the most likely to be published thanks to their favourable results.

Conversely, the 5 data points with medium SE on the left side of the plot, just outside the lower 95% CI bound, indicate studies with effect estimates considerably lower than the overall weighted mean Dice.

The 6 data points inside the 95% CI bounds (within the red diagonals) refer to studies whose mean Dice scores were consistent with overall findings.

To visualize the exhaustive set of plots we produced, please go to [Appendix.A.B](#AppendixA).

## Figure 16 – Egger’s Test Summary Statistics

A screenshot of a computer

Description automatically generated

Image showing the summary statistics related to Egger’s test, which was conducted as part of the meta-analysis in order to verify the findings form the funnel plot in terms of publication bias.

The slope indicates the change in SE associated with a one-unit increase in Dice. Since it is negative (-0.033), it suggests that higher Dice is associated with lower SE, with statistical significance (p=0<0.05), thus confirming the previously observed publication bias.

# Ethical Approval

We believe that our pilot analysis does not require ethical approval for the following reasons:

* The re-use of ISLES 2015 SISS is explicitly allowed for the intended purpose of this project [13]
* The approval from “SMIR” (SICAS Medical Image Repository) for ISLES 2015 SISS dataset as part of the account creation process, before getting access to the actual datasets
* ISLES 2015 SISS dataset was officially approved by the local ethics committee: “The local ethics committee approved their release under Az.14-256A” [13]

# References

1. Subudhi, A., Dash, P., Mohapatra, M., Tan, R.-S., Acharya, U. R., & Sabut, S. (2022). Application of Machine Learning Techniques for Characterization of Ischemic Stroke with MRI Images: A Review. Diagnostics, 12(10).
2. 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.
3. Lee, E.-J., Kim, Y.-H., Kim, N., & Kang, D.-W. (2017). Deep into the Brain: Artificial Intelligence in Stroke Imaging. Journal of Stroke, 19(3), 277–285.
4. 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).
5. 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.
6. 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.
7. Balakrishnan, R., del C. Valdés Hernández, M., & Farrall, A. J. (2021). Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data – A systematic review. Computerized Medical Imaging and Graphics, 88, 101867.
8. Mainali, S., Darsie, M. E., & Smetana, K. S. (2021). Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. Frontiers in Neurology, 12.
9. 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.
10. 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).
11. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. Journal of Digital Imaging, 30(4), 449–459.
12. Offersen, C. M., Sørensen, J., Sheng, K., Carlsen, J. F., Langkilde, A. R., Pai, A., Truelsen, T. C., & Nielsen, M. B. (2023). Artificial Intelligence for Automated DWI/FLAIR Mismatch Assessment on Magnetic Resonance Imaging in Stroke: A Systematic Review. Diagnostics, 13(12).
13. 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.
14. 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).
15. Saka, O., McGuire, A., & Wolfe, C. (2008). Cost of stroke in the United Kingdom. Age and Ageing, 38(1), 27–32.
16. Abbasi, H., Orouskhani, M., Asgari, S., & Zadeh, S. S. (2023). Automatic brain ischemic stroke segmentation with deep learning: A review. Neuroscience Informatics, 3(4), 100145.
17. 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.
18. 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.
19. Lo, E. H. (2008). A new penumbra: transitioning from injury into repair after stroke. Nature Medicine, 14(5), 497–500.
20. Moskowitz, M. A., Lo, E. H., & Iadecola, C. (2010). The Science of Stroke: Mechanisms in Search of Treatments. Neuron, 67(2), 181–198.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. Veit, A., Wilber, M., & Belongie, S. (2016). Residual Networks Behave Like Ensembles of Relatively Shallow Networks.
28. 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.
29. 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.
30. 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.
31. 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).
32. Caceres, P. (2020). Introduction to Neural Network Models of Cognition (NNMOC). https://com-cog-book.github.io/com-cog-book/features/cov-net.html.
33. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. Lee, A.-R., 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.
47. 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.
48. Diganta, M. (2020). Attention Mechanisms in Computer Vision: CBAM. Https://Blog.Paperspace.Com/Attention-Mechanisms-in-Computer-Vision-Cbam.
49. Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. https://arxiv.org/abs/1709.01507.
50. 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.
51. Takyar, A. (2021). How Attention Mechanism’s Selective Focus Fuels Breakthroughs in AI. https://Www.Leewayhertz.Com/Attention-Mechanism/.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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).
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. Pavlakis, S. G., Hirtz, D. G., & deVeber, G. (2006). Pediatric Stroke: Opportunities and Challenges in Planning Clinical Trials. Pediatric Neurology, 34(6), 433–435.
65. 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.
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. 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.
68. Iglovikov, V., & Shvets, A. (2018). TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. https://arxiv.org/abs/1801.05746.
69. 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.
70. 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.
71. 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.
72. Yu, F., & Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. https://arxiv.org/abs/1511.07122.
73. 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.
74. 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.
75. 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.
76. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(56), 1929–1958.
77. Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. BMJ, 327(7414), 557–560.
78. 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/.
79. 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.
80. 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.
81. 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.
82. Malik, M., Chong, B., Fernandez, J., Shim, V., Kasabov, N. K., & Wang, A. (2024). Stroke Lesion Segmentation and Deep Learning: A Comprehensive Review. Bioengineering, 11(1), 86.
83. Karthik, R., Menaka, R., Johnson, A., & Anand, S. (2020). Neuroimaging and deep learning for brain stroke detection - A review of recent advancements and future prospects. Computer Methods and Programs in Biomedicine, 197, 105728.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. 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.
95. Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? https://arxiv.org/abs/1703.04977.
96. 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.
97. 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.
98. 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.
99. 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.
100. Zaharchuk, G. (2020, March 19). Imaging Collaterals in Acute Stroke (iCAS) (iCAS). Https://Clinicaltrials.Gov/Study/NCT02225730.
101. 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.
102. 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.
103. Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for Simplicity: The All Convolutional Net. https://arxiv.org/abs/1412.6806.
104. Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. Pp. 818–833.
105. 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.
106. Hedges, L. (1985). Statistical Methods for Meta-Analysis. Elsevier.
107. Schmidt, R. L., & Factor, R. E. (2013). Understanding sources of bias in diagnostic accuracy studies. Archives of Pathology & Laboratory Medicine, 137(4), 558–565.
108. Sitburana, O., & Koroshetz, W. J. (2005). Magnetic resonance imaging: Implication in acute ischemic stroke management. Current Atherosclerosis Reports, 7(4), 305–312.
109. 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.
110. 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.
111. Wardlaw, J. M., & Farrall, A. J. (2004). Diagnosis of stroke on neuroimaging. BMJ, 328(7441), 655–656.
112. 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.
113. 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.
114. 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.
115. 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.
116. Liu, L., Wu, F.-X., & Wang, J. (2019). Efficient multi-kernel DCNN with pixel dropout for stroke MRI segmentation. Neurocomputing, 350, 117–127.
117. 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.
118. 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.
119. Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M. C., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index1. Academic Radiology, 11(2), 178–189.
120. 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.
121. 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.
122. 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.
123. 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.
124. 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.
125. 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.