Improved Segmentation and Detection Sensitivity of Diffusio n-weighted Stroke Lesions with Synthetically Enhanced Deep Learning

Nov 05, Journal Club, tAILab

Hyunjae Jeong (tAILab, CCIDS)

Yonsei University, Medical Life Systems Information Center (TAIL Lab)
Severance Hospital, Center for Clinical Imaging Data Science (CCIDS)
Severance Hospital, Radiology

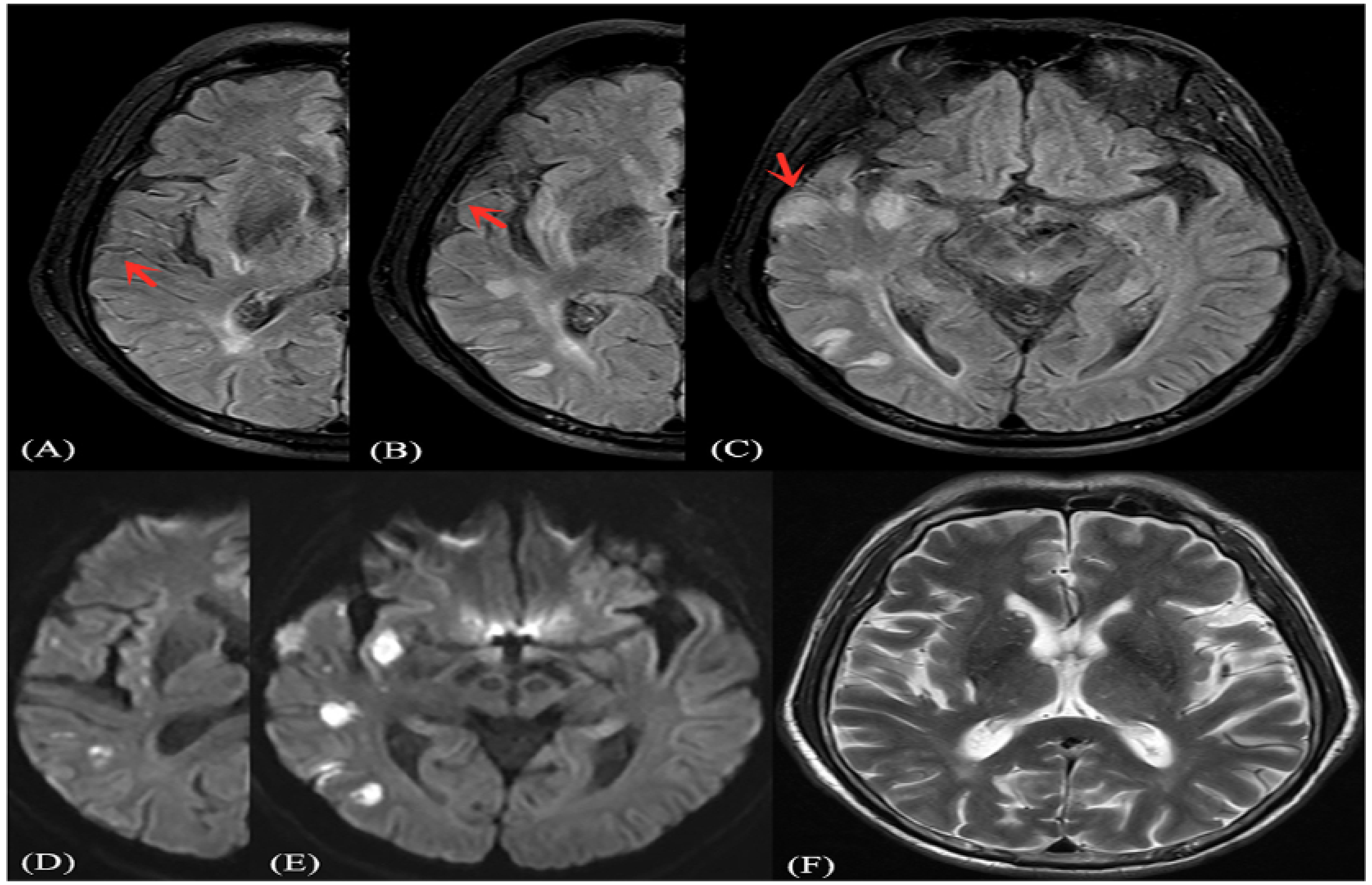
의료영상데이터사이언스센터
Center for Clinical Imaging Data Science

tAILab.

01 Introduction

Segmentation and Detection Sensitivity

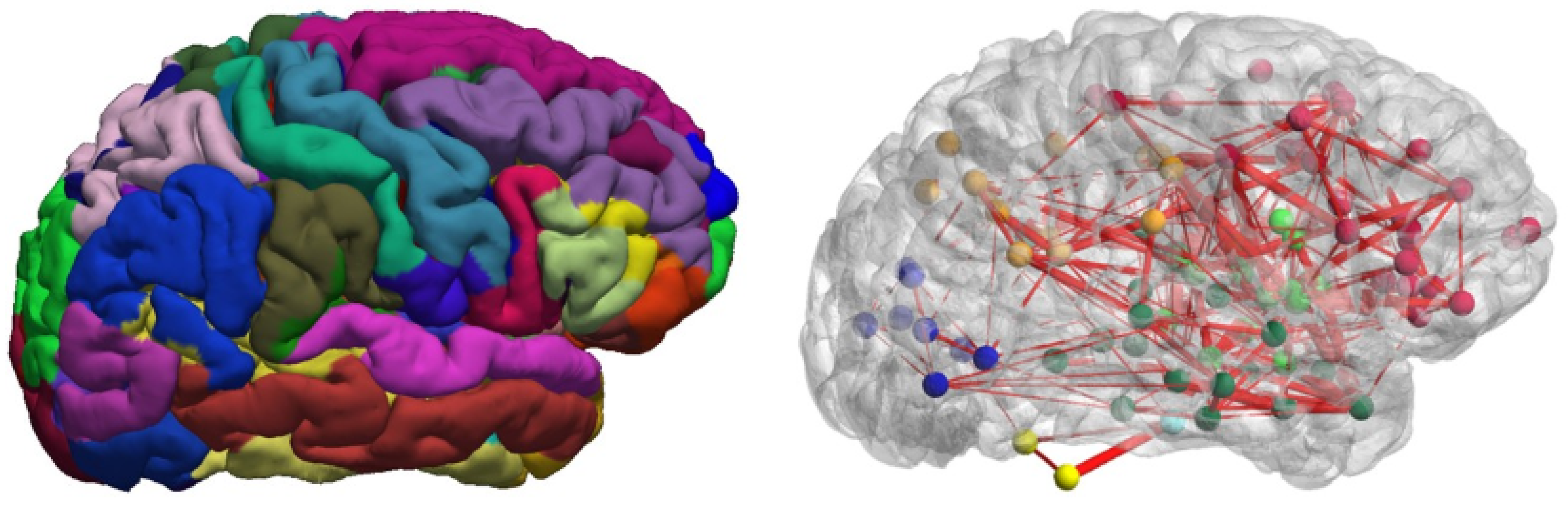
- Detection of stroke lesions can be challenging, and mistakes in diagnosis can lead to delays in therapy with pot ential harmful consequences for the patient.
- Missing a small embolic stroke can lead to patient discharge without adequate clarification of the source of stroke
- Progress in machine learning for medical image analysis is hampered by the small size of the databases availab le to train models.



← Small Lesions of Ischemic Stroke

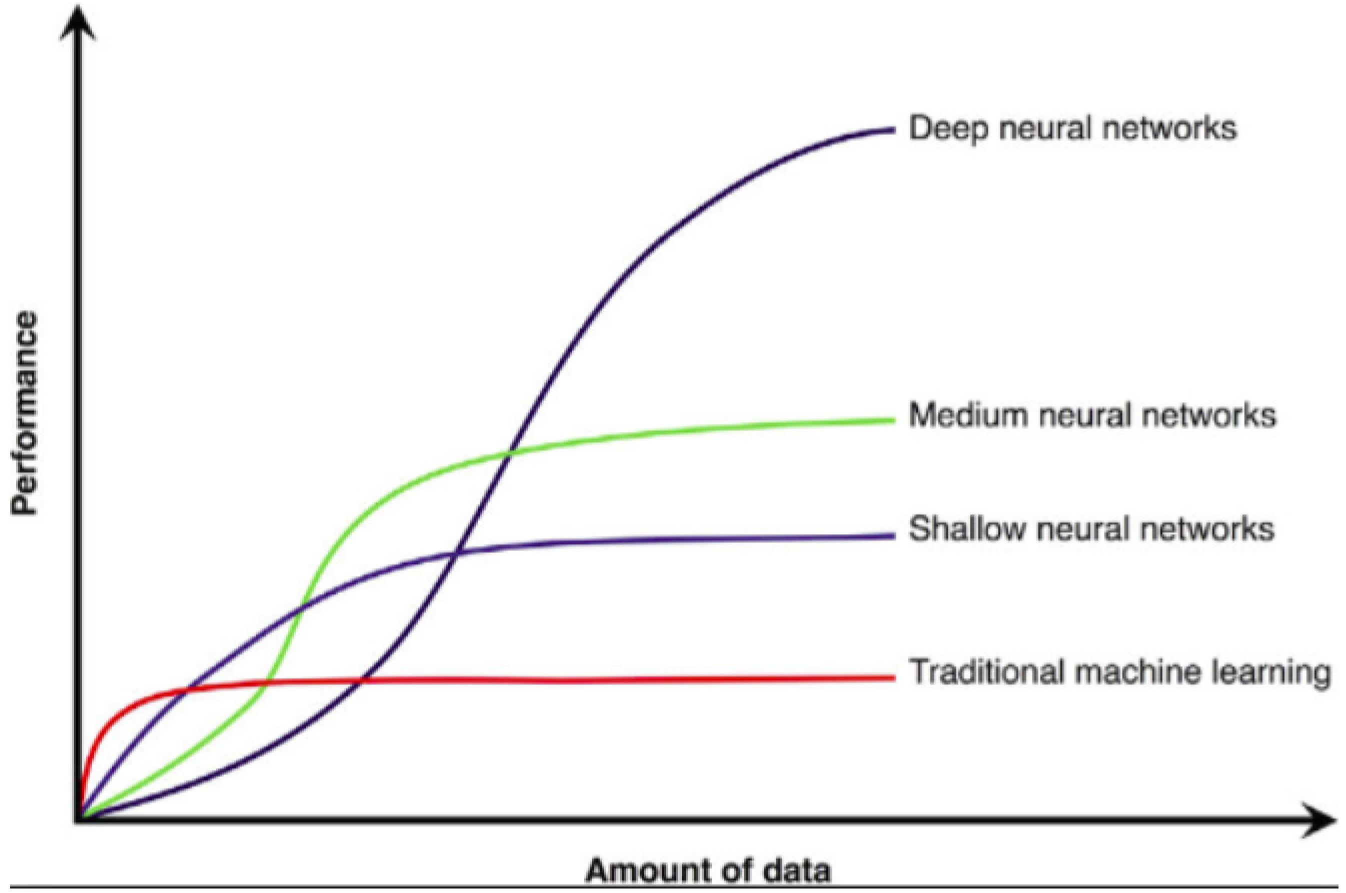
Lack of Data & Synthetic Stroke Image Generation

- The labeling of radiologic images is tedious, requires a high level of expertise, is error prone, and can be very subjective
- Because of the various problems (e.g. ethical and privacy issues, medicolegal aspects, patient protection, and v arious technical challenges, laborsome), There is little hope for a substantial improvement in the number of available labeled medical images for training
- We hypothesized that the performance of a network trained on a synthetically enriched dataset would be better than a network trained on clinical data alone.



(a) (b)

Brain region's highly structured through spatial layout



Contribution

- To generate a large number of DW volumes with synthetic stroke lesions from a stroke database and a no rmal database to independently train a generic three-dimensional (3D) U-Net
- To compare the diagnostic accuracy in lesion segmentation and detection of these networks on a separat e test set comprising volumes in patients with clinical suspicion of stroke.

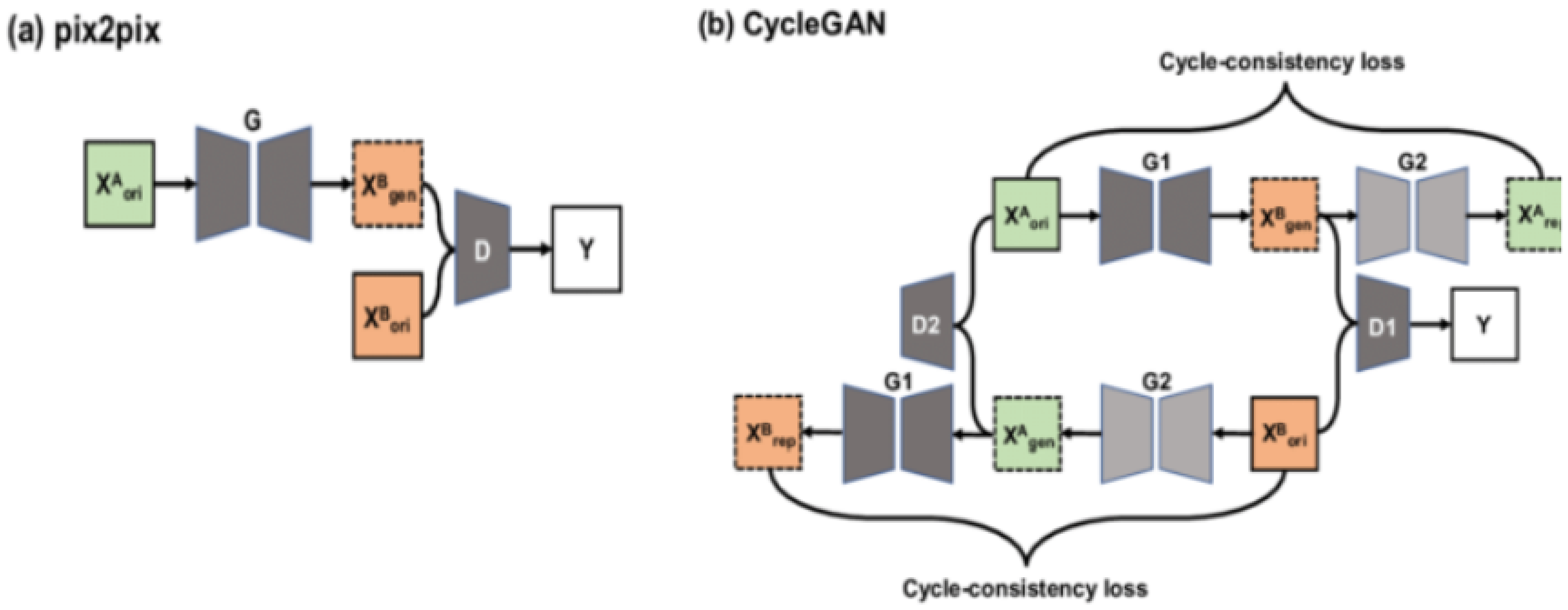
↓

- The authors developed a robust algorithm capable of generating a large quantity of diffusion-weighted M R images with realistic synthetic stroke lesions
- Combination of human-labeled clinical stroke volumes and synthetic stroke volumes significantly outperf ormed the same 3D U-Net trained only on the human-labeled clinical stroke volumes

Related Work - Synthetic Image Generation

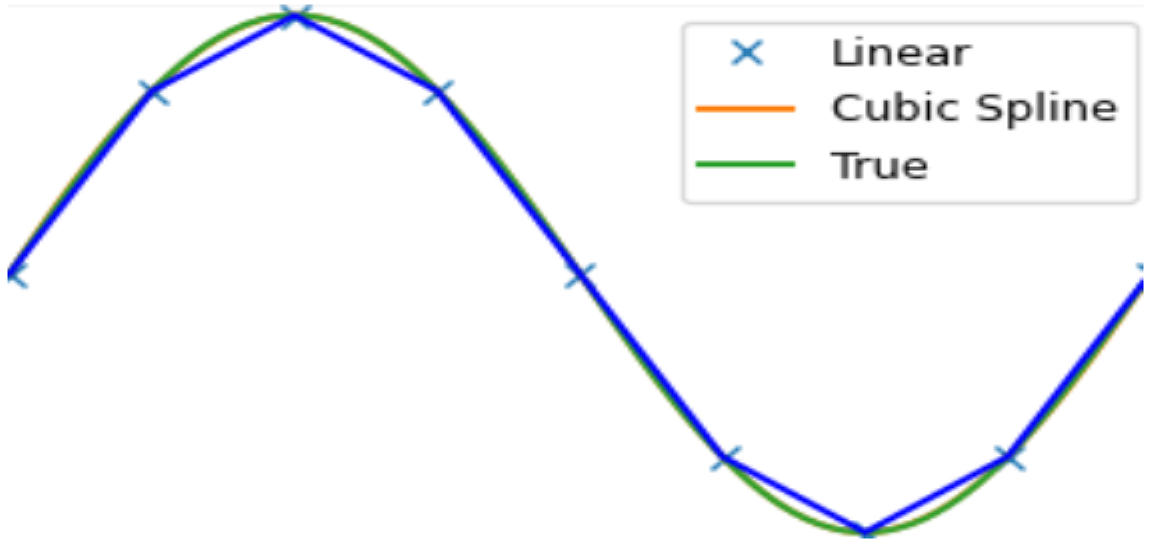
<GAN - Image Generation>

Supervised Image-to-Image Translations are more data-dependent, requiring well-annotated paired training samples.
Unsupervised image-to-image translation problem is inherently ill-posed. Additional constraints are needed



<Spline Interpolation>

Spline interpolation is a form of interpolation where the interpolant is a special type of piecewise polynomial called a spline.
In this Journal, They use Deformable B-spline transformation and third-order spline resampling



$$\mathbf{S}(t) = \sum_{i=0}^{m-n-2} \mathbf{P}_i b_{i,n}(t)\ ,\ t \in [t_n, t_{m-n-1}].$$

Original Datasets & New Generated Datasets

- Original Datasets : Stroke database and Normal database.
  The stroke database included DW positive 962 cases + 513 without DW-positive stroke lesions,
  The normal database included 2027 cases
- New Generated Datasets (CS2DB, CS40DB, S2DB) : stroke database and normal database were randomly selected (with replacement). Then, The stroke volume was coregistered to the normal image using a deformable B-spline transformation and third-order spline resampling using ANTS.

Normal Database

N = 2027

Stroke Database

DW+ N = 449

DW- N = 513

Spline Transformation

CS2DB

80% of Positive DW+

2,000 Synthetic Stroke

CS40DB

80% of Positive DW+

40,000 Synthetic Stroke

S2DB

2,000 Synthetic Stroke

1. Normalization

normalized by dividing by the maximum after clipping.

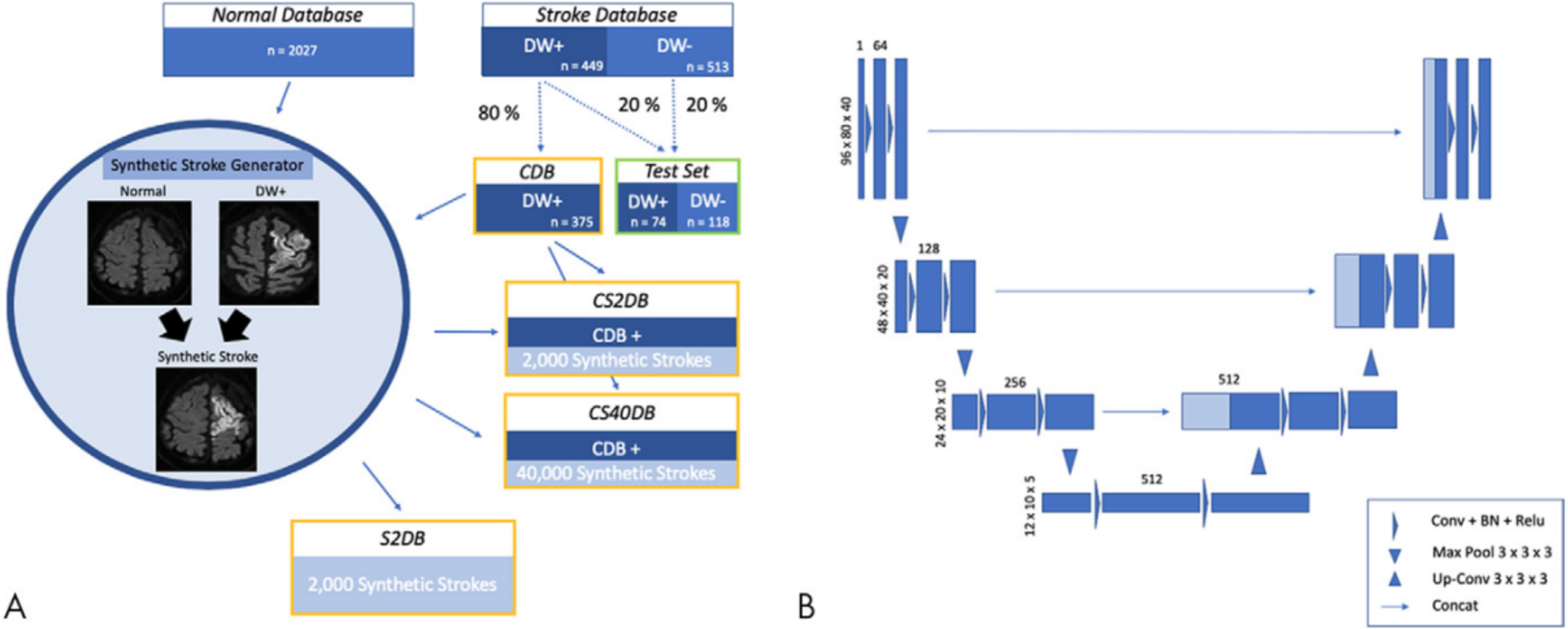
2. Data Augmentation

Randomly left-right flip, translation [-0.02, 0.02], zoom [0.95, 1.05], stretch [0.95, 1.05], shear [-0.03, 0.03]

3. Hyperparameter

Optimizer = Adam, Epoch = 1000, Loss Function = Cross Entropy + Weighted Dice Coef

Final Proposed Framework



Evaluation Metrics

- Quantitative Analysis, Statistical Analysis

Quantitative Analysis : Mean Dice, Median Dice, Min-Max Range Dice, Stroke Volume
Statistical Analysis : P-Value vs S2DB, CS2DB, CS40DB

- Detection Analysis

Sensitivity $$\textbf{Sensitivity} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Negatives}}$$

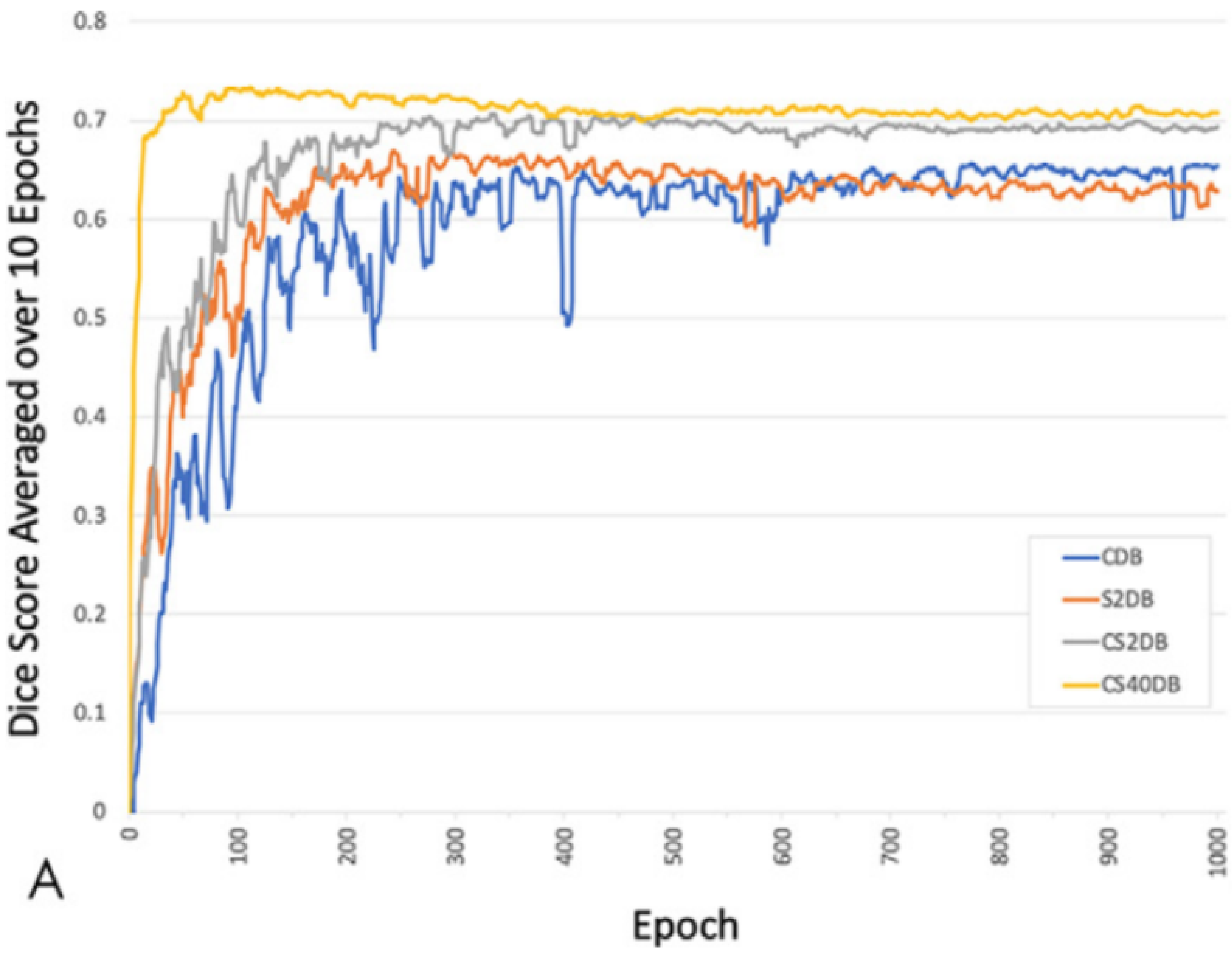
Specificity $$\textbf{Specificity} = \frac{\textbf{True Negatives}}{\textbf{True Negatives} + \textbf{False Positives}}$$

PPV Positive predictive value $$\frac{Sensitivity \times Prevalence}{Sensitivity \times Prevalence \ + \ (1 - Specificity) \times (1 - Prevalence)}$$

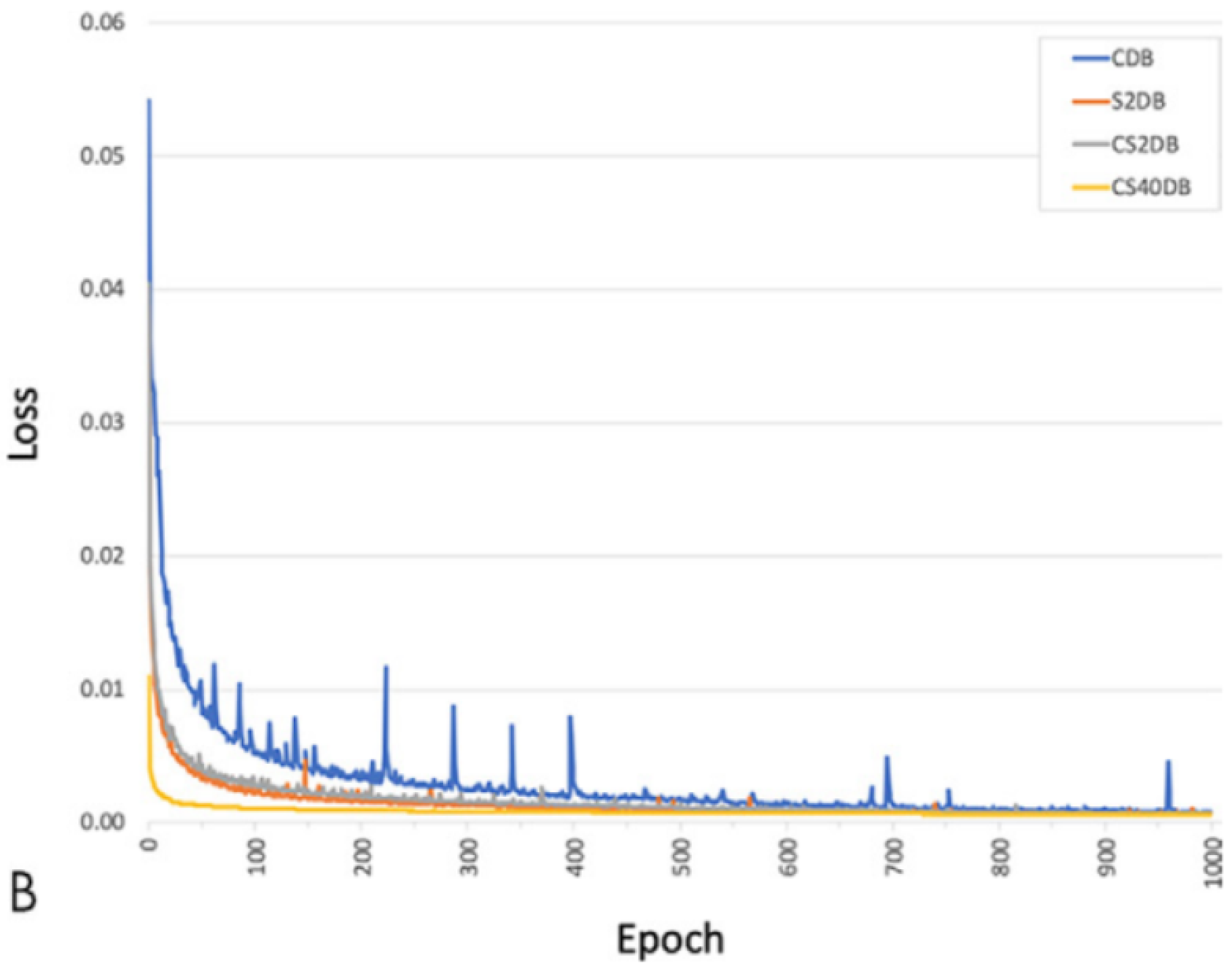
NPV Negative predictive value $$\frac{Specificity \times (1 - Prevalence)}{Specificity \times (1 - Prevalence) \ + \ (1 - Sensitivity) \times Prevalence}$$

03 Experiments

Results



A



B

03 Experiments

Results

CS2DB

80% of Positive DW+

2,000 Synthetic Stroke

CS40DB

80% of Positive DW+

40,000 Synthetic Stroke

S2DB

2,000 Synthetic Stroke

Table 1: Overview of Dice Coefficient of the Test Set for the Models Trained with the Clinical and Synthetic Databases

Database	Mean Dice	Median Dice	Min–Max Range Dice	Error in Stroke Volume [mL]	*P* Value vs S2DB	*P* Value vs CS2DB	*P* Value vs CS40DB
CDB	0.618 ± 0.077	0.639	0.005–0.670	5 ± 15	$2\cdot10^{-8}$	$2\cdot10^{-30}$	$2\cdot10^{-42}$
S2DB	0.653 ± 0.036	0.661	0.338–0.680	-2 ± 3	...	$5\cdot10^{-32}$	$7\cdot10^{-62}$
CS2DB	0.695 ± 0.027	0.700	0.457–0.720	1 ± 1	...	...	$2\cdot10^{-18}$
CS40DB	0.721 ± 0.010	0.725	0.677–0.743	1 ± 0	...	...	...

Note.—Dice coefficient of the test set (n = 192) achieved in the "trained range" for each model. Data are mean ± standard deviation, median, minimum (Min) and maximum (Max) range of the Dice score as well as the error in stroke volume as mean ± standard deviation of the trained range from epoch 250 to 450, except for the CS40DB, which used a trained range from epoch 100 to 300. *P* values represent the paired two-tailed Student *t* test of the Dice coefficient of the respective model in the trained range. CDB = database of 375 human-labeled clinical stroke cases, CS2DB = database of 375 human-labeled clinical stroke cases plus 2000 synthetic cases, CS40DB = database of 375 human-labeled clinical stroke cases and 40 000 synthetic cases, S2DB = database of 2000 synthetic cases.

03 Experiments

Results

CS2DB

80% of Positive DW+

2,000 Synthetic Stroke

CS40DB

80% of Positive DW+

40,000 Synthetic Stroke

S2DB

2,000 Synthetic Stroke

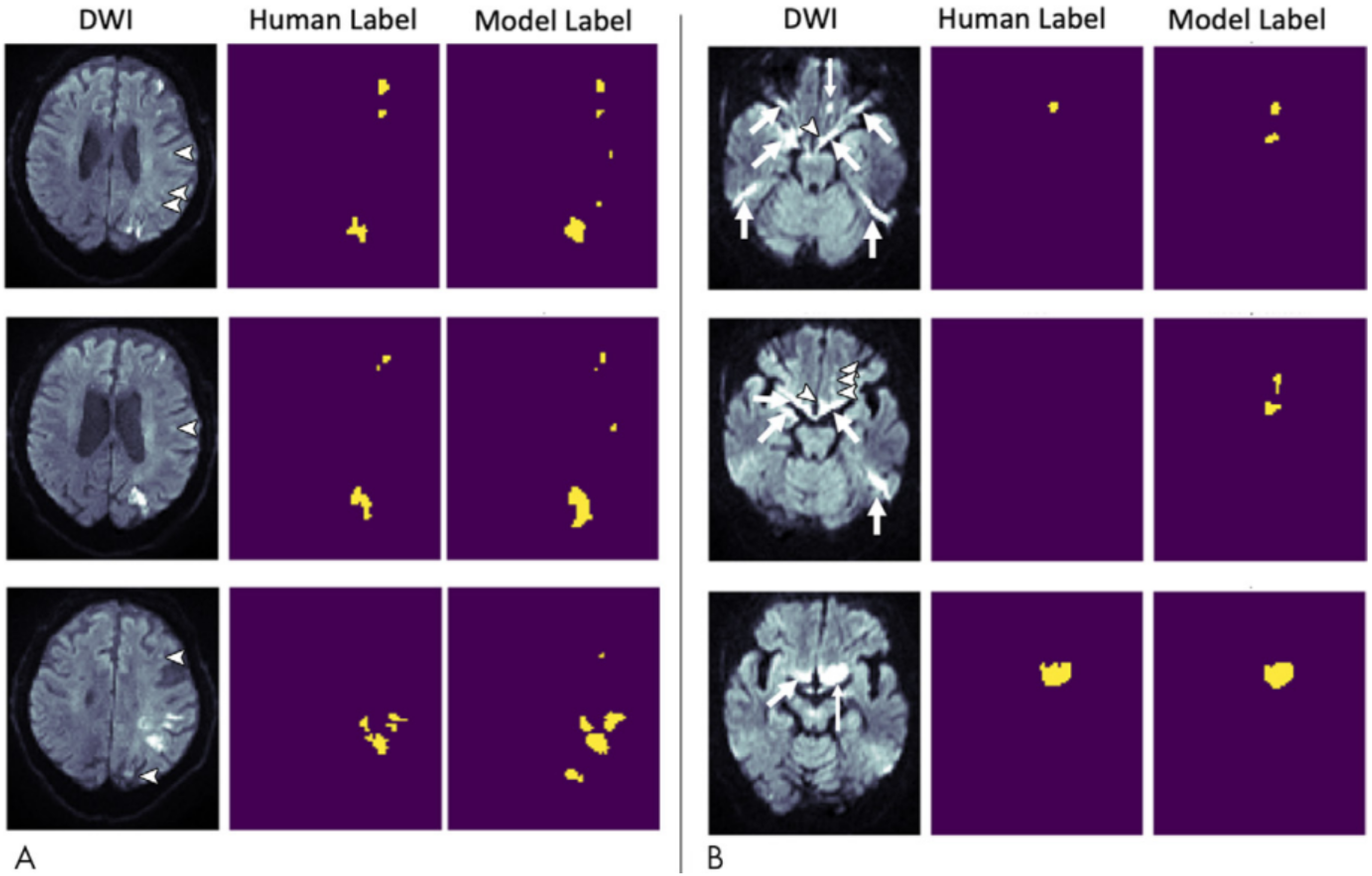
Table 2: Detailed Stroke Lesion Detection Analysis

Analysis Type	Epoch	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
CDB	423	85 (82, 87)	48 (44, 51)	64 (61, 67)	74 (69, 78)
S2DB	296	77 (73, 79)	68 (65, 72)	73 (70, 76)	72 (69, 76)
CS2DB	274	80 (77, 83)	76 (72, 79)	79 (76, 82)	78 (74, 81)
CS40DB	154	91 (89, 93)	75 (72, 78)	80 (77, 83)	88 (85, 91)
Human reader 1	...	78 (75, 81)	92 (90, 94)	92 (89, 94)	79 (76, 82)
Human reader 2	...	79 (76, 82)	89 (86, 91)	89 (86, 91)	79 (76, 82)
Human reader 3 (reference standard)	...	84 (81, 87)	96 (94, 98)	96 (94, 97)	85 (82, 87)

Note.—Detailed stroke lesion detection analysis in a random subset (n = 80 patients) of the test set, with 95% confidence interval in parentheses. CDB = database of 375 human-labeled clinical stroke cases, CS2DB = database of 375 human-labeled clinical stroke cases plus 2000 synthetic cases, CS40DB = database of 375 human-labeled clinical stroke cases and 40 000 synthetic cases, S2DB = database of 2000 synthetic cases, NPV = negative predictive value, PPV = positive predictive value.

03 Experiments

Results



Conclusion

- A combination of human-labeled clinical stroke images and the synthetic stroke images significantly outperformed the models trained on the human-labeled clinical stroke images alone and rivaled human abilities in terms of sensitivity.
- This work demonstrates that the shortcoming of having relatively few medical images for training can be overcome by producing realistic synthetic images.
- Further work should investigate using generative methods, such as GAN or VAE and if the final performance on segmentation and lesion detection could be further increased.

Limitation

- There was a significant difference between the age of patients in the stroke database and the normal database.
- Compared with the humans could be accounted for by the tiredness and time limitations of the human readers
- A large number of parameters was selected by relying upon experience and good engineering principles without direct evaluation