IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 30, 2020, accepted May 9, 2020, date of publication May 19, 2020, date of current version June 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995632

Automatic Segmentation of Stroke Lesions in Non-Contrast Computed Tomography Datasets With Convolutional Neural Networks

ANUP TULADHAR[1,2,*], SERENA SCHIMERT[1,3,*], DEEPTHI RAJASHEKAR[1,2], HELGE C. KNIEP[4], JENS FIEHLER[4], AND NILS D. FORKERT[1,2,3,5]

[1]Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada
[2]Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada
[3]Department of Clinical Neurosciences, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada
[4]Department of Diagnostic and Interventional Neuroradiology, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany
[5]Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB T2N 4N1, Canada

Corresponding author: Nils D. Forkert (nils.forkert@ucalgary.ca)

*Anup Tuladhar and Serena Schimert are co-first authors.

This work was supported by the Heart and Stroke Foundation of Canada under Grant G-17-0018368. The work of Anup Tuladhar was supported in part by the T. Chen Fong Fellowship in Medical Imaging Science, and in part by the Natural Sciences and Engineering Researsch Council of Canada (NSERC) Postdoctoral Fellowship. The work of Serena Schimert was supported by the University of Calgary Program for Undergraduate Research Experience (PURE) award.

ABSTRACT Non-contrast computed tomography (NCCT) is commonly used for volumetric follow-up assessment of ischemic strokes. However, manual lesion segmentation is time-consuming and subject to high inter-observer variability. The aim of this study was to develop and establish a baseline convolutional neural network (CNN) model for automatic NCCT lesion segmentation. A total of 252 multi-center clinical NCCT datasets, acquired from 22 centers, and corresponding manual segmentations were used to train (204 datasets) and validate (48 datasets) a 3D multi-scale CNN model for lesion segmentation. Post-processing methods were implemented to improve the CNN-based lesion segmentations. The final CNN model and post-processing method was evaluated using 39 out-of-distribution holdout test datasets, acquired at seven centers that did not contribute to the training or validation datasets. Each test image was segmented by two or three neuroradiologists. The Dice similarity coefficient (DSC) and predicted lesion volumes were used to evaluate the segmentations. The CNN model achieved a mean DSC score of 0.47 on the validation NCCT datasets. Post-processing significantly improved the DSC to 0.50 ($P < 0.01$). On the holdout test set, the CNN model achieved a mean DSC score of 0.42, which was also significantly improved to 0.45 ($P < 0.05$) by post-processing. Importantly, the automatically segmented lesion volumes were not significantly different from the lesion volumes determined by the expert observers ($P > 0.05$) and showed excellent agreement with manual lesion segmentation volumes (intraclass correlation coefficient, ICC = 0.88). The proposed CNN model can automatically and reliably segment ischemic stroke lesions in clinical NCCT datasets. Post-processing techniques can further improve accuracy. As the model was trained and evaluated on datasets from multiple centers, it is broadly applicable and is publicly available.

INDEX TERMS Artificial neural networks, brain, computed tomography, computer-assisted image analysis, convolutional neural networks, deep learning, machine learning, stroke.

I. INTRODUCTION

Non-contrast computed tomography (NCCT) is the most common imaging modality for volumetric assessment of stroke lesions [1], [2]. Manual lesion segmentation in NCCT images is time consuming and associated with high inter-observer variability. Semi-automatic lesion segmentation tools have been developed [3], [4] but still require human interaction, which could introduce a bias, while previous work on automatic NCCT lesion segmentation is limited [5], [6].

Deep convolutional neural networks (CNNs) have shown superior performance for various segmentation tasks in medical imaging because of their ability to learn complex patterns and relationships in the data [7]. Convolutional kernels in CNNs enable the learning of non-localized spatial relationships. The use of multi-scale features and three-dimensional (3D) kernels [8] allows an automated seg-

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

mentation algorithm to take advantage of the spatial contiguity of stroke lesions while maintaining localized context. However, for stroke lesion segmentation, these methods have only been applied to magnetic resonance imaging (MRI) [9]–[11] or computed tomography perfusion and angiography datasets [12], [13]. To date, multi-scale 3D CNNs for a fully automatic stroke lesion segmentation have not been evaluated in NCCT datasets, despite its common application in stroke imaging [1], [2].

Thus, the aim of this work was to train and evaluate a multi-scale 3D CNN model for stroke lesion segmentation in follow-up NCCT datasets. For further improvement of the CNN segmentations, post-processing methods were investigated. We tested the model's generalizability by evaluating it on an out-of-distribution holdout test set, using multi-center datasets acquired at entirely different centers that did not contribute to the training and validation sets.

II. METHODS

A. DATASETS

A total of 291 clinical follow-up NCCT datasets from the ESCAPE (252 datasets) [14] and ERASER (39 datasets) [15] multi-center trials were used. These datasets were acquired across 29 centers and corresponding manual segmentations were available. Expert observers manually segmented patient lesions in three orthogonal planes simultaneously using ITK-SNAP [16]. The in-slice resolution ranged from 0.355 to 0.637 mm, the slice thickness ranged from 1.00 to 10.0 mm, and the number of slices ranged from 10 to 141.

Approval and informed consent for the datasets from the two trials was approved by the respective ethics board at each site contributing to the two trials. All datasets used in this retrospective secondary study were made available after complete anonymization. Thus, additional ethics approval and informed consent were not required.

1) TRAINING AND VALIDATION SETS

The ESCAPE datasets collected from 22 centers were used for training and validation of the 3D CNN-based lesion segmentation model. Out of the 252 ESCAPE datasets, 204 were used for training and 48 were used for validation.

2) OUT-OF-DISTRIBUTION HOLDOUT TEST SET

The 39 ERASER datasets used for the out-of-distribution holdout test set were collected from seven centers that did not contribute to the training or validation sets. Experienced neuroradiologists manually segmented the test datasets. Each example was segmented by two (19 datasets) or three (20 datasets) observers with multiple years of dedicated experience in stroke imaging. By using out-of-distribution datasets that were segmented by multiple expert observers, this completely independent test set provides a stronger estimate of the model's generalization performance.

B. NCCT SCAN PRE-PROCESSING

As NCCT images were acquired from multiple centers with differing scanners and imaging protocols, training a CNN directly on NCCT images without pre-processing resulted in very poor performance on the training set (data not shown). Thus, the datasets were pre-processed to ensure consistency between NCCT images collected from different centers.

First, the bone structures were removed from each dataset, retaining only the brain tissue in the images. To remove the bone structures, which have high Hounsfield values, a six-step procedure following the approach described by Muschelli *et al.* [17] was performed in a slice-wise manner. This approach was implemented using the Insight Segmentation and Registration Toolkit (ITK) [18]. Briefly described, a Gaussian filter with a variance of 4 pixels is used to smooth each slice. In the next step, the intensities are thresholded between 0 and 100 Hounsfield units, which removes most of the artifacts from bone and other high-intensity tissue. After this, a circular structural element with a radius of 1 pixel is used to erode the resulting segmentation. Subsequently, the largest connected component in each slice is extracted and a circular structural element with a radius of 1 pixel is used to dilate this component in order to create a brain mask for the slice. After performing the these three steps in each slice, the masks from each slice are combined into a final mask for the entire volume and any holes in this final mask are filled using the ''Voting Binary Hole Filling Image Filter'' in ITK. Finally, the images were thresholded again between 0 and 100 Hounsfield units to remove the remaining high-intensity tissue artifacts resulting from the morphological erosion and dilation. The images are then normalized to zero mean and unit variance to account for potential differences in scanner tube potential and different reconstruction algorithms. All images in the training, validation, and holdout test datasets underwent the same pre-processing procedure.

C. CNN ARCHITECTURE

The CNN used in this work is based on the DeepMedic model proposed by Kamnitsas *et al.* [8] and modified for NCCT stroke lesion segmentation. The network parameters were optimized with cross-validation. We used a total of 11 layers. The first eight layers consist of three parallel convolutional pathways for processing the images at multiple scales. The multi-scale pathways were created by using down-sampled versions of the NCCT images (by factors of $3\times$ and $5\times$) as inputs to the parallel convolutional pathways, in addition to the original image. Each parallel pathway has eight convolutional layers consisting of 30, 30, 40, 40, 40, 40, 50, and 50 feature maps and uses convolutional kernels of size $3 \times 3 \times 3$. Additionally, residual skip connections between layers two and four, between layers four and six, and between layers six and eight are used in each parallel pathway. The ninth layer combines the three multi-scale pathways together by using the concatenated outputs from layer eight of each parallel pathway. Layer nine uses $3 \times 3 \times 3$ convolutional kernels and has 250 feature maps. Layer ten is a fully-connected convolutional layer with $1 \times 1 \times 1$ convolutional kernels and 250 feature maps. Additionally, a residual skip connection between layers eight and ten was used. The final softmax

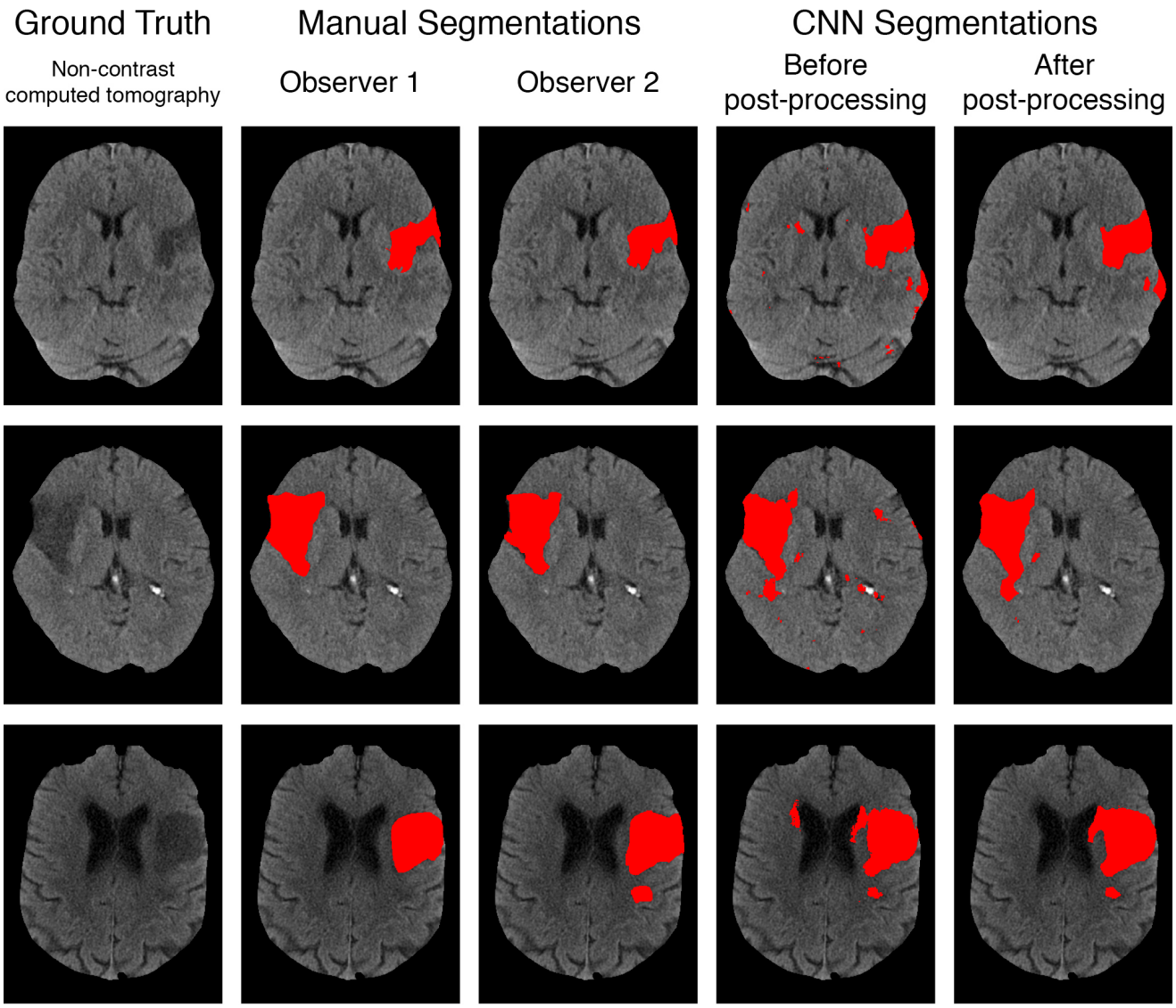

FIGURE 1. Examples of manual and CNN-based segmentations from the independent holdout test dataset.

classification layer, layer eleven, produces the lesion probability maps. A threshold of >0.5 is used to binarize the probability map to a final lesion segmentation.

D. CNN TRAINING

All CNN model training was performed in Python 2.7 on Compute Canada and Calcul Quebec computing clusters. The DeepMedic framework (v.0.7.3), available from https://github.com/deepmedic/deepmedic, was used for model training. The DeepMedic framework performs model training on image segments extracted from the original image, rather than the entire image. In this work, segments of $37 \times 37 \times 37$ were used. The network was trained for 35 epochs with a batch size of 10. Each epoch was divided into 20 sub-epochs, within which 1000 image segments were extracted and used for model training. An initial learning rate of 0.001, which decreases through training using a polynomial decay function, was employed. Root mean square propagation was used as the optimizer. L1 and L2 regularizations of 10^{-6} and 10^{-4} were used, respectively. Data augmentation consisted of mirroring along the sagittal axis. The CNN model achieved a mean Dice similarity coefficient (DSC) of 0.52 in the training set evaluated by 10-fold cross-validation.

The DSC scores and lesion volumes for the automatic segmentations of the validation and holdout test sets were obtained using a single CNN model that was trained on the entire training data.

E. POST-PROCESSING OF CNN SEGMENTATIONS

The CNN-based binary lesion segmentations were post-processed to further improve the segmentation accuracy (Fig. 1). The implemented post-processing consists of a connected component analysis to exclude small lesion components, most likely caused by noise artifacts, and an automatic hole-filling approach. Post-processing was performed using

the ITK toolkit. Using a connected components analysis, components smaller than an empirically determined cut-off were removed. The exception was in segmentations where the largest connected component was smaller than the cut-off value, in which case no cutoff was applied. Afterwards, a hole-filling algorithm was used to fill gaps within the segmentation.

The validation dataset was used to estimate the optimum minimum object size cut-off and the hole-filling kernel radius. The minimum object size cut-off was optimized first, by varying the cut-off range from 0.3 cm^3 to 2.5 cm^3. The cut-off that maximized the DSC was 1.5 cm^3 and was used for post-processing.

Using this minimum object size cut-off, the hole-filling radius was optimized next using values of 2, 3, 5, 7, and 10 voxels. As hole-filling causes the segmented lesion volumes to grow, and subsequently increased the error in lesion volume estimates, both the DSC and lesion volume error were considered when choosing the optimal value. More precisely, the DSC was maximized while the lesion volume error was minimized. The optimal radius was found to be 3 voxels and was used for post-processing.

F. SEGMENTATION EVALUATION

The DSC was used as the primary outcome measurement for evaluation of the automatic lesion segmentations. The DSC measures the overlap between two segmentations and is defined between 0 and 1, where 1 indicates perfect consensus. The DSC is calculated as:

$$DSC = (2^*|A \cap B|)/(|A| + |B|)$$

where A and B are two binary segmentations of the same dataset.

The DSC scores in the validation set, for which only a single manual segmentation was available for each dataset, were computed by comparing the manual segmentations to the CNN-based methods (CNN or CNN + post-processing).

The DSC scores for CNN-based methods in the holdout test set, for which two or three manual lesion segmentations were available for each dataset, were computed independently for each observer (*e.g.* CNN vs. observer A, CNN vs. observer B, for two observers) and averaged together using the arithmetic mean. This average value was then used as the DSC score for the CNN-based methods (CNN or CNN + post-processing) vs. Observers.

The inter-observer DSC for the holdout test set was calculated for pairs of observers. In examples with three observers, three pair-wise DSC scores (*e.g.* observer A vs. observer B, observer A vs. observer C, observer B vs. observer C) were calculated and averaged together using the arithmetic mean to obtain a single, average, inter-observer DSC score.

Lesion volumes were calculated by multiplying the number of lesion voxels in the binary segmentation mask by the volume of each voxel. As multiple observers segmented the holdout test set, an average lesion volume estimate was obtained for each example. This was calculated as the arithmetic mean of the lesion volumes segmented by the individual observers.

Intra-class correlation coefficients (ICC) were used to assess inter-rater reliability in lesion volume estimates [19]. ICC was calculated for absolute agreement between observers for manual lesion volume estimates, and for absolute agreement between manual lesion volume estimates (average of observers) and automated lesion volume estimates (CNN or CNN with post-processing) [20]. ICCs above 0.75 were considered as excellent inter-rater reliability, following the guidelines established by the American Psychological Association [21].

G. STATISTICS

Results are reported as mean ± standard deviation (SD) or median [interquartile range] as appropriate. The Wilcoxon signed-rank test or Friedman test with Dunn's multiple comparison post-hoc correction was used for comparisons. Correlation was quantified using Spearman's rank correlation. Statistical significance was set as $P<0.05$. Significance in figures is denoted by *($P < 0.05$) and **($P < 0.01$). All statistical analyses were performed using Graphpad Prism 8.4.

III. RESULTS

The median lesion volumes for the training and validation sets were 40.4 [14.1–96.3] cm^3 and 41.5 [20.0–107.1] cm^3, respectively. As the out-of-distribution holdout test set was segmented by multiple observers, the manual segmentation lesion volume for each example was defined as the average volume calculated across observers. It was not possible to ensure the volume distribution of the holdout test set is similar to the training and validation sets. This is because the test set was drawn from an entirely different multi-center trial and completely independent of the training and validation sets. The median lesion volume for the test set was 20.9 [9.7–63.7] cm^3, which is considerably lower compered to the training and validation sets.

A. VOXEL-WISE AGREEMENT

The CNN model achieved a mean DSC score of 0.47 ± 0.22 in the validation set. Post-processing significantly improved the DSC to 0.50 ± 0.23 (Wilcoxon signed-rank test, $P < 0.01$; Fig. 2).

The model's generalizability was assessed using the out-of-distribution holdout test set (Fig. 1). These datasets were manually segmented by multiple independent observers and had an inter-observer DSC score of 0.73 ± 0.13 (Fig. 3). The CNN lesion segmentations had a DSC score of 0.42 ± 0.25 compared to manual segmentations (average of observers), which was lower than the inter-observer DSC (Friedman test with Dunn's multiple comparisons, $P < 0.01$). Post-processing of CNN-based segmentations significantly improved the DSC to 0.45 ± 0.26 (Friedman test with Dunn's multiple comparisons, $P < 0.05$).

The decreased DSC score on the out-of-distribution holdout test data may be partly attributable to an abundance

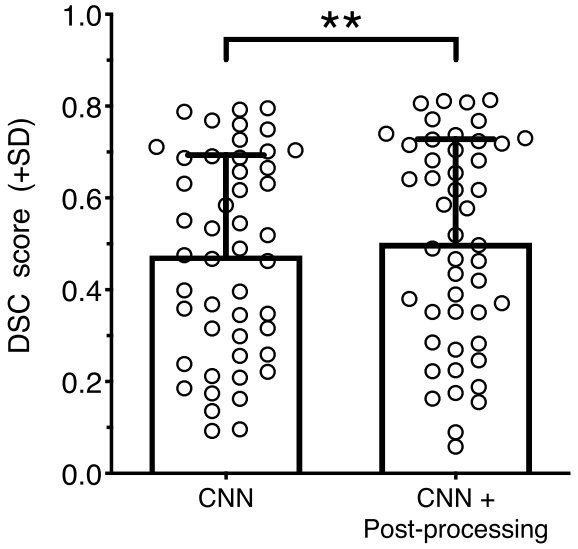

FIGURE 2. Evaluation of DSC score comparing manual segmentations with the trained CNN model and post-processing on the validation set. Bar plots express the mean + standard deviation (SD).

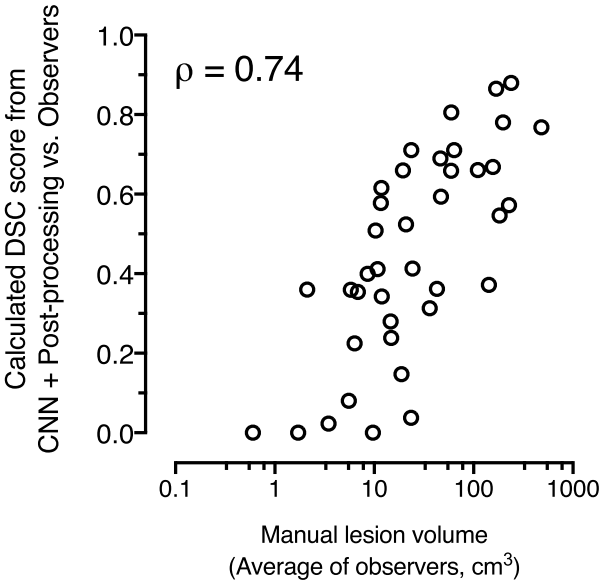

FIGURE 4. Correlation between test set DSC scores from CNN-based segmentations after post-processing with lesion volumes, indicating that automated segmentations perform better on larger lesions.

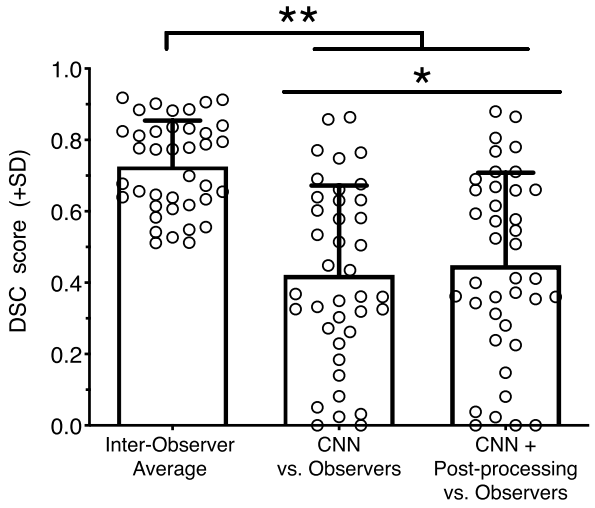

FIGURE 3. Comparison of DSC scores from inter-observer agreement and CNN-based automated segmentations vs. observer segmentations on the out-of-distribution test set segmented by multiple observers. Bar plots express the mean + standard deviation (SD).

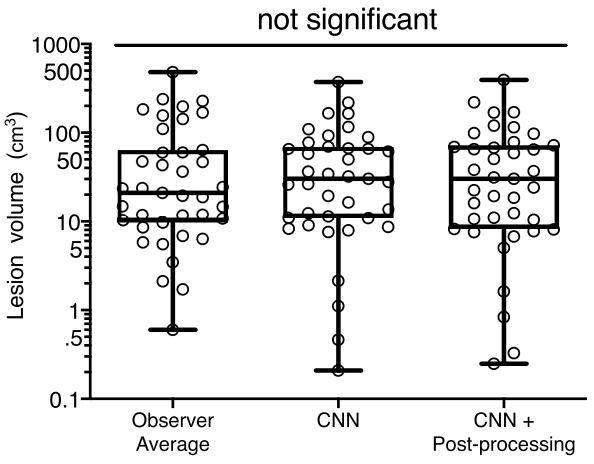

FIGURE 5. Calculated lesion volumes in the out-of-distribution test set for CNN segmentations and CNN segmentations after post-processing were not significantly different from manual segmentations. Box plots express the median and inter-quartile range.

of smaller lesions in the test set, as the manual vs. automated DSC scores showed a strong positive correlation with manually segmented lesion volumes (before post-processing: Spearman's $\rho = 0.77$, $P < 0.01$; after post-processing: Spearman's $\rho = 0.74$, $P < 0.01$, Fig. 4). Indeed, even the inter-observer agreement was lower on smaller lesions and showed a strong positive correlation with lesion volume (Spearman's $\rho = 0.68$, $P < 0.01$).

B. LESION VOLUME ESTIMATES

Median lesion volume estimates for the holdout test datasets calculated from CNN lesion segmentations before (30.1[10.9–68.9] cm^3) and after (30.2 [8.2–72.2] cm^3) post-processing were not significantly different from lesion volumes measured by manual segmentations (20.9 [9.7–63.7] cm^3) (Friedman test with Dunn's multiple comparisons, $P > 0.05$, Fig. 5). Bland-Altman analysis reflected a tendency for the model to over-predict lesion volumes, though this bias was minimal (Fig. 6).

Importantly, lesion volume estimates from CNN segmentations showed excellent agreement with manual segmentations. The ICC for CNN lesion segmentations was 0.86, which was further improved by post-processing to 0.88 (Fig. 7). The agreement between observers for manual segmentations was lower, with an ICC of 0.80.

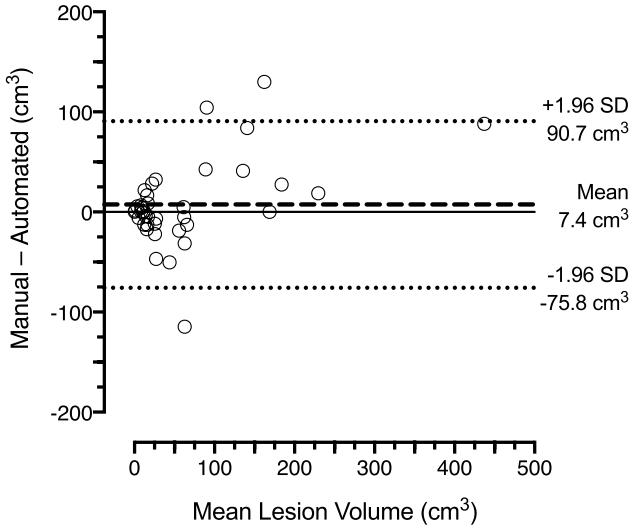

FIGURE 6. Bland-Altman agreement analysis of test set lesion volumes from manual segmentations and CNN segmentations with post-processing, showing minimal bias. SD: standard deviation.

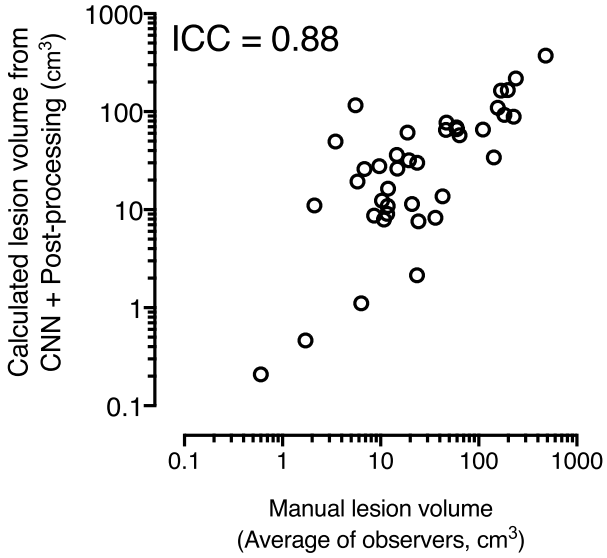

FIGURE 7. Excellent agreement between calculated lesion volumes from automatic and manual lesion segmentations in the out-of-distribution test set. ICC: Intraclass correlation coefficient.

IV. DISCUSSION

In this study, we developed an automatic method for clinical NCCT stroke lesion segmentation using a 3D multi-scale CNN. Volumetric assessments from the automatic CNN-based method were in excellent agreement with multiple neuroradiologists. We used an out-of-distribution test set, which was completely independent from the training and validation data, in order to obtain a reliable estimate of the method's generalizability and establish a reproducible baseline for automated NCCT stroke lesion segmentation with CNNs. As NCCT is a standard imaging procedure available in most stroke centers for follow-up assessment, a generalizable automatic lesion segmentation pipeline for this modality is of high demand.

CNN models for automatic follow-up lesion segmentation have primarily been developed and investigated for MRI [8]–[11], [22]–[24]. Though reported DSC scores for lesion segmentation in MRI are typically higher (0.67–0.79) than seen in our study, the proposed method is nevertheless a promising approach for NCCT segmentation. Lesion segmentation in NCCT is more challenging compared to typical MRI follow-up sequences such as diffusion-weighted MRI [25], as the ischemic changes in NCCT images are more subtle.

To prevent over-fitting of the CNN model to a specific dataset and imaging protocol, it was trained using multi-center NCCT datasets. Testing the model on out-of-distribution datasets with multiple manual expert observer segmentations from seven completely independent centers revealed that our model generalizes well. Decreases in voxel-wise agreement in the holdout test set, compared to the validation set, may be attributable to multiple factors. First, as previously mentioned, the test dataset came from independent centers that did not contribute to the training or validation datasets. This stands in contrast to the often used method of creating a test set by sampling from the same data source as the training set, which increases the risk of over-fitting the model to a specific data distribution and inflating model performance. Evaluating the model on test data from independent centers as done in this study may provide a more reliable and reproducible estimate of model's generalizability.

Second, the test dataset was segmented by multiple observers who did not contribute segmentations to the training or validation sets. The inter-observer voxel-wise agreement was variable, with an inter-observer DSC of 0.73. This variability in the ground truth segmentations can be expected to impact the voxel-wise agreement of automatically extracted segmentations. However, evaluating the model using segmentations from multiple observers who did not contribute to the training set segmentations may further strengthen the estimate of model's generalizability.

Other recent studies on automatic NCCT segmentation of stroke lesions have used single-scale CNN models [5], [6], reporting mean DSC scores (0.54–0.57) and volumetric ICCs (0.88) comparable to ours (DSC = 0.45, ICC = 0.88). However, these studies only evaluated their models on test data that belonged to the same distribution as their training set and were only compared against single reference segmentations [5], [6]. How these methods generalize to datasets outside their training distribution and against multiple expert observers, important for evaluating their broader applicability, was not evaluated in detail in the original studies. As the developed models are not publicly available, it was not possible to evaluate them in this study. In contrast to this, our model was tested on out-of-distribution datasets from seven completely independent centers with segmentations from multiple expert observers, which demonstrated that our model generalizes well.

Finally, the test dataset contained a greater number of small lesions. Due to the higher surface area to volume ratio, a high voxel-wise agreement for small lesions is harder to achieve. This was also reflected in the fact that manual segmentations also showed less agreement for smaller lesions. A similar result was seen for CNN segmentations of NCCT images by Barros *et al.*, where the DSC scores were lower for subtle injuries with smaller stroke lesions (0.37) [5].

Though the voxel-wise agreement of the CNN-based segmentations was inferior to the inter-observer agreement, the corresponding lesion volumes were not significantly different from manual segmentations and had excellent agreement with them. The agreement between automatically and manually segmented lesion volumes was higher than the inter-observer agreement between manual segmentations. This suggests a potential application of the model for consistent volumetric assessment of follow-up lesions in multi-center studies. Providing consistent results is an advantage of automatic segmentation algorithms, thereby reducing variability between sites or studies.

To improve the CNN-based lesion segmentations, simple post-processing techniques were used to correct for the speckly nature of CNN segmentations. However, more sophisticated analyses such as hidden Markov random fields may achieve further improvements [26].

Though promising, this preliminary study has some limitations. First, the CNN architecture used was originally designed for MRI images [8]. While this DeepMedic model establishes a baseline for NCCT segmentation with multi-scale 3D CNNs, future work may investigate modifications such as applying the well-known U-Net [27] for improved multi-scale segmentations. Second, segmentations of smaller lesions need to be improved. Future work may investigate training models specifically for segmenting small lesions, whether through CNN architecture modifications or using a small lesion training dataset. As clinical datasets may be difficult to acquire, data augmentation methods such as generative adversarial networks may be explored [28]. Third, the training of deep learning models is stochastic in nature. Future work may investigate ensembling multiple models trained on subsets of the training data, as this may allow averaging out deficiencies in single models [5], [10], [24]. Addressing these limitations may further improve CNN-based lesion segmentations in NCCT datasets.

V. CONCLUSION

This study demonstrated the successful use of a CNN-based automated method for follow-up stroke lesion segmentation in clinical NCCT datasets. This lays the foundation for developing more advanced automatic lesion analysis tools for NCCT images and can contribute toward consistent and high-throughput analysis of large multi-center studies.

To facilitate further development of NCCT lesion segmentation methods and to provide a baseline for future evaluations, the trained CNN model is publicly available [29].

DATA AVAILABILITY

The trained model is available for use from http://dx.doi.org/10.21227/jps9-0b57 [29].

ACKNOWLEDGMENT

The authors would like to thank Drs. Michael D. Hill, Mayank Goyal, and Andrew M. Demchuk for sharing data from the ESCAPE trial (NCT01778335) that formed part of this study's training and validation sets.

DISCLOSURES

The Authors declare that there is no conflict of interest.

REFERENCES

[1] M. Wintermark, M. Luby, N. M. Bornstein, A. Demchuk, J. Fiehler, K. Kudo, K. R. Lees, D. S. Liebeskind, P. Michel, R. G. Nogueira, M. W. Parsons, M. Sasaki, J. M. Wardlaw, O. Wu, W. Zhang, G. Zhu, and S. J. Warach, "International survey of acute stroke imaging used to make revascularization treatment decisions," *Int. J. Stroke*, vol. 10, no. 5, pp. 759–762, Jul. 2015.

[2] J. Schröder and G. Thomalla, "A critical review of Alberta stroke program early CT score for evaluation of acute stroke imaging," *Frontiers Neurol.*, vol. 7, p. 245, Jan. 2017.

[3] H. Kuang, B. K. Menon, and W. Qiu, "Semi-automated infarct segmentation from follow-up noncontrast CT scans in patients with acute ischemic stroke," *Med. Phys.*, vol. 15, p. 283, Jul. 2019.

[4] H. Kuang, B. K. Menon, and W. Qiu, "Segmenting hemorrhagic and ischemic infarct simultaneously from follow-up non-contrast CT images in patients with acute ischemic stroke," *IEEE Access*, vol. 7, pp. 39842–39851, 2019.

[5] R. S. Barros *et al.*, "Automatic segmentation of cerebral infarcts in follow-up computed tomography images with convolutional neural networks," *J. NeuroIntervent. Surg.*, 2019, doi: 10.1136/neurintsurg-2019-015471.

[6] T. Fuchigami, S. Akahori, T. Okatani, and Y. Li, "A hyperacute stroke segmentation method using 3D U-net integrated with physicians' knowledge for NCCT," in *Proc. SPIE, Med. Imag., Comput.-Aided Diagnosis*, vol. 11314, Mar. 2020, Art. no. 113140G, doi: 10.1117/12.2549176.

[7] G. Zaharchuk, E. Gong, M. Wintermark, D. Rubin, and C. P. Langlotz, "Deep learning in neuroradiology," *Amer. J. Neuroradiology*, vol. 39, no. 10, pp. 1776–1784, Oct. 2018.

[8] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[9] L. Chen, P. Bentley, and D. Rueckert, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *NeuroImage Clin.*, vol. 15, pp. 633–643, Jan. 2017.

[10] O. Wu *et al.*, "Big data approaches to phenotyping acute ischemic stroke using automated lesion segmentation of multi-center magnetic resonance imaging data," *Stroke*, vol. 50, no. 7, pp. 1734–1741, Jul. 2019.

[11] L. Liu, F.-X. Wu, and J. Wang, "Efficient multi-kernel DCNN with pixel dropout for stroke MRI segmentation," *Neurocomputing*, vol. 350, pp. 117–127, Jul. 2019.

[12] O. Öman, T. Mäkelä, E. Salli, S. Savolainen, and M. Kangasniemi, "3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke," *Eur. Radiol. Exp.*, vol. 3, no. 1, pp. 8–11, Feb. 2019.

[13] A. S. Kasasbeh, S. Christensen, M. W. Parsons, B. Campbell, G. W. Albers, and M. G. Lansberg, "Artificial neural network computer tomography perfusion prediction of ischemic core," *Stroke*, vol. 50, no. 6, pp. 1578–1581, Jun. 2019.

[14] A. M. Demchuk *et al.*, "Endovascular treatment for small core and anterior circulation proximal occlusion with emphasis on minimizing CT to recanalization times (ESCAPE) trial: Methodology," *Int. J. Stroke*, vol. 10, no. 3, pp. 429–438, Apr. 2015.

[15] J. Fiehler, G. Thomalla, M. Bernhardt, H. Kniep, A. Berlis, F. Dorn, B. Eckert, A. Kemmling, S. Langner, L. Remonda, W. Reith, S. Rohde, M. Möhlenbruch, M. Bendszus, N. D. Forkert, and S. Gellissen, "ERASER," *Stroke*, vol. 50, no. 5, pp. 1275–1278, May 2019.

[16] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, Jul. 2006.
[17] J. Muschelli, N. L. Ullman, W. A. Mould, P. Vespa, D. F. Hanley, and C. M. Crainiceanu, "Validated automatic brain extraction of head CT images," *NeuroImage*, vol. 114, pp. 379–385, Jul. 2015.
[18] T. S. Yoo, M. J. Ackerman, W. E. Lorensen, W. Schroeder, V. Chalana, S. Aylward, D. Metaxas, and R. Whitaker, "Engineering and algorithm design for an image processing Api: A technical report on ITK–the Insight Toolkit," *Stud. Health Technol. Inf.*, vol. 85, pp. 586–592, 2002.
[19] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability.," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.
[20] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropractic Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016.
[21] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.," *Psychol. Assessment*, vol. 6, no. 4, pp. 284–290, Dec. 1994.
[22] O. Maier, C. Schröder, N. D. Forkert, T. Martinetz, and H. Handels, "Classifiers for ischemic stroke lesion segmentation: A comparison study," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0145118.
[23] R. Zhang, L. Zhao, W. Lou, J. M. Abrigo, V. C. T. Mok, W. C. W. Chu, D. Wang, and L. Shi, "Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional DenseNets," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2149–2160, Sep. 2018.
[24] S. Winzeck, S. J. T. Mocking, R. Bezerra, M. J. R. J. Bouts, E. C. McIntosh, I. Diwan, P. Garg, A. Chutinet, W. T. Kimberly, W. A. Copen, P. W. Schaefer, H. Ay, A. B. Singhal, K. Kamnitsas, B. Glocker, A. G. Sorensen, and O. Wu, "Ensemble of convolutional neural networks improves automated segmentation of acute ischemic lesions using multiparametric diffusion-weighted MRI," *Amer. J. Neuroradiol.*, vol. 40, no. 6, pp. 938–945, Jun. 2019.
[25] J. B. Fiebach, P. D. Schellinger, O. Jansen, M. Meyer, P. Wilde, J. Bender, P. Schramm, E. Jüttler, J. Oehler, M. Hartmann, S. Hähnel, M. Knauth, W. Hacke, and K. Sartor, "CT and diffusion-weighted MR imaging in randomized order: Diffusion-weighted imaging results in higher accuracy and lower interrater variability in the diagnosis of hyperacute ischemic stroke," *Stroke*, vol. 33, no. 9, pp. 2206–2210, Sep. 2002.
[26] N. K. Subbanna, D. Rajashekar, B. Cheng, G. Thomalla, J. Fiehler, T. Arbel, and N. D. Forkert, "Stroke lesion segmentation in FLAIR MRI datasets using customized Markov random fields," *Frontiers Neurol.*, vol. 10, p. 541, May 2019.
[27] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense, volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, vol. 9901, no. 1. Cham, Switzerland: Springer, 2016, pp. 424–432.
[28] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," *Sci. Rep.*, vol. 9, no. 1, pp. 16884-1–16884-9, Nov. 2019.
[29] A. Tuladhar, S. Schimert, D. Rajashekar, H. Kniep, J. Fiehler, and N. Forkert, "Automatic segmentation of stroke lesions in non-contrast computed tomography datasets with convolutional neural networks," *IEEE Dataport*, 2020. Accessed: May 20, 2020, doi: 10.21227/jps9-0b57.

ANUP TULADHAR was born in Kathmandu, Nepal, in 1987. He received the Bachelor of Applied Science (B.A.Sc.) degree in chemical engineering from the University of Waterloo, Waterloo, ON, Canada, in 2011, and the Ph.D. degree in biomedical engineering from the University of Toronto, Toronto, ON, Canada, in 2018.

From 2005 to 2011, he was an Undergraduate Research Assistant with various labs as part of the University of Waterloo's Co-Op Program, including the Defence Research and Development Canada, in 2008, the Nippon Telegraph and Telecommunication's Sensory and Emotion Research Group, from 2009 to 2010, and the Bio-Acoustic MEMS in Medicine Laboratory, Harvard, in 2010. He is currently a Postdoctoral Fellow in machine learning and medical imaging with the University of Calgary, Calgary, AB, Canada. His research interests include machine learning for medical informatics, data privacy, distributed machine learning, and biological and artificial neural networks.

Dr. Tuladhar was a recipient of the T. Chen Fong Postdoctoral Fellowship in Medical Imaging Science and the Natural Science and Engineering Research Council (NSERC) Postdoctoral Fellowship.

SERENA SCHIMERT was born in Vancouver, Canada, in 1998. She is currently pursuing the combined B.S. degree in neuroscience and computer science with the University of Calgary, Calgary, Canada.

In summer months of 2017, 2018, and 2019, she completed studentships in various research labs associated with the Hotchkiss Brain Institute, Calgary. Her research interests include novel applications of machine learning networks in medical imaging processing, and more recently refining the prediction of patient outcomes following stroke, with emphasis on language deficit.

Ms. Schimert was a recipient of the Program for Undergraduate Research Experience (PURE) Award from the University of Calgary.

DEEPTHI RAJASHEKAR was born in Karnataka, India, in 1991. She received the B.Eng. degree from PES University, India, in 2013, majoring in information science and pattern recognition, and the M.Sc. degree in computer science from Dalhousie University, Canada, in 2017. She is currently pursuing the Ph.D. degree with the University of Calgary.

Her research focus was user behavior modeling using machine learning and artificial intelligence techniques for cybersecurity applications. In 2014, she was a Research Assistant with the Center for Pattern Recognition, PES University. Her research interests include medical image analysis, lesion-deficit mapping in stroke patients, deep-learning, and computer-aided decision support systems using machine learning methods.

Ms. Rajashekar was a recipient of the Harley N. Hotchkiss Doctoral Graduate Award funded by the Hotchkiss Brain Institute, University of Calgary.

HELGE C. KNIEP was born in Hamburg, Germany, in 1985. He received the Dipl.-Ing degree in mechanical engineering and industrial engineering from the Technical University of Hamburg-Harburg, Germany, in 2011.

From 2008 to 2009, he has served as Visiting Scholar of mechanical engineering with the University of California at Berkeley. He was with McKinsey & Company, from 2012 to 2014, ultimately as Senior Associate. Since 2014, he is affiliated with the Department of Diagnostic and Interventional Neuroradiology, University Medical Center Hamburg-Eppendorf. His research interest includes application of artificial intelligence in the context of conception and conduction of multi-center medical studies. He is particularly interested in image analysis and advanced study methodology in interventional neuroradiology trials.

Mr. Kniep is a member of the Deutsche Röntgengesellschaft and the American Society of Neuroradiology. His awards and honors include Full Scholarships of the German Academic Scholarship Foundation (Studienstiftung des Deutschen Volkes) and the German Academic Exchange Service (DAAD).

JENS FIEHLER was born in Thuringia, Germany, in 1972. He received the Dr.Med. (M.D.) degree in pathophysiology from the University of Jena, Germany, in 2001, and the Habilitation (Ph.D.) degree in radiology from the University Medical Center Hamburg-Eppendorf, Hamburg, Germany, in 2005. He completed board examinations in radiology, in 2005, neuroradiology, in 2006, and quality management in medicine, in 2007.

From 2008 to 2016, he was a Visiting Professor with the Department of Neuroradiology, Oxford University, U.K. Since 2009, he has been a Professor and the Chairman of the Department of Diagnostic and Interventional Neuroradiology, University Medical Center Hamburg-Eppendorf, Germany. He has authored or coauthored more than 400 peer-reviewed articles. His research interest includes the conception and conduct of multi-center studies in cerebrovascular diseases. He is particularly interested in imaging analysis and advanced study methodology in Interventional Neuroradiology trials.

Prof. Fiehler aims to foster and support high-quality research in European Interventional Neuroradiology, as an Associate Editor for the *Journal of Neurointerventional Surgery* (JNIS) and *Clinical Neuroradiology* (CNR).

Nils D. Forkert received the German Diploma degree in computer science from the University of Hamburg, in 2009, the master's degree in medical physics from the Technical University of Kaiserslautern, in 2012, and the Ph.D. degree in computer science from the University of Hamburg, in 2013.

He completed a Postdoctoral Fellowship at Stanford University before joining the University of Calgary, in 2014. He is currently an Assistant Professor and the Canada Research Chair with the Departments of Radiology and Clinical Neurosciences, University of Calgary. He is also an Imaging and Machine Learning Scientist who develops new image processing methods, predictive algorithms, and software tools for the analysis of medical data. This includes the extraction of clinically relevant imaging parameters and biomarkers describing the morphology and function of organs with the aim of supporting clinical studies and preclinical research as well as developing computer-aided diagnosis and patient-specific prediction models using machine learning based on multi-modal medical data.