TBI Lesion Segmentation in Head CT: Impact of Preprocessing and Data Augmentation

Miguel Monteiro[1], Konstantinos Kamnitsas[1], Enzo Ferrante[2], Francois Mathieu[3], Steven McDonagh[1], Sam Cook[3], Susan Stevenson[3], Tilak Das[3], Aneesh Khetani[3], Tom Newman[3], Fred Zeiler[3], Richard Digby[3], Jonathan P. Coles[3], Daniel Rueckert[1], David K. Menon[3], Virginia F.J. Newcombe[3], and Ben Glocker[1]

[1] Biomedical Image Analysis Group, Imperial College London, UK
[2] sinc(i), FICH-Universidad Nacional del Litoral, CONICET, Argentina
[3] Division of Anaesthesia, Department of Medicine, University of Cambridge, UK

Abstract. Automatic segmentation of lesions in head CT provides key information for patient management, prognosis and disease monitoring. Despite its clinical importance, method development has mostly focused on multi-parametric MRI. Analysis of the brain in CT is challenging due to limited soft tissue contrast and its mono-modal nature. We study the under-explored problem of fine-grained CT segmentation of multiple lesion types (core, blood, oedema) in traumatic brain injury (TBI). We observe that preprocessing and data augmentation choices greatly impact the segmentation accuracy of a neural network, yet these factors are rarely thoroughly assessed in prior work. We design an empirical study that extensively evaluates the impact of different data preprocessing and augmentation methods. We show that these choices can have an impact of up to 18% DSC. We conclude that resampling to isotropic resolution yields improved performance, skull-stripping can be replaced by using the right intensity window, and affine-to-atlas registration is not necessary if we use sufficient spatial augmentation. Since both skull-stripping and affine-to-atlas registration are susceptible to failure, we recommend their alternatives to be used in practice. We believe this is the first work to report results for fine-grained multi-class segmentation of TBI in CT. Our findings may inform further research in this under-explored yet clinically important task of automatic head CT lesion segmentation.

1 Introduction

Traumatic brain injury (TBI) is a pathology that alters brain function caused by trauma to the head [1]. TBI is a leading cause of death and disability worldwide with heavy socio-economic consequences [2]. Computed tomography (CT) allows rapid assessment of brain pathology, ensuring patients who require urgent intervention receive appropriate care [3]. Its low acquisition time allows for rapid diagnosis, quick intervention, and safe application to trauma and unconscious patients. Research on automatic segmentation of TBI lesions in magnetic resonance imaging (MRI) [4,5] has shown promising results. However, MRI is usually

reserved for imaging in the post-acute phase of brain injury or as a research tool. Since CT is routinely used in clinical care and CT voxel intensities are approximately calibrated in Hounsfield units (HUs) across different scanners, effective computational analysis of CT has the potential for greater generalisation and clinical impact than MRI. Prior work on automatic analysis of pathology in head CT is limited, mostly focusing on image-level detection of abnormalities [6,7], feature extraction for outcome prediction [8], or image-level classification [9] instead of voxel-wise semantic segmentation. Previous works on segmentation employ level-sets for the segmentation of specific haemorrhages and haematomas [10,11]. Recently, [12] applied deep learning for binary segmentation of contusion core grouped with haematomas. We present the first multi-class segmentation of contusion core, blood (haemorrhages and haematomas), and oedema. This task is important for patient management and a better understanding of TBI.

State-of-the-art automatic segmentation relies on convolutional neural networks (CNNs) [13]. These models are effective in many biomedical imaging tasks [14]. Neural networks are theoretically capable of approximating any function [15]. This result commonly translates in the expectation that networks are able to extract any necessary pattern from the input data during training. As a result, attention is mainly focused on further development of network architectures, while disregarding other parts of the system, such as data preprocessing. Contrary to popular belief that networks generalise well by learning high-level abstractions of the data, they tend to learn low-level regularities in the input [16]. In practice, the learned representations are largely dependent on a stochastic, greedy and non-convex optimisation process. We argue that appropriate data preprocessing and augmentation can be as important as architectural choices. Preprocessing can remove useless information from the input and help training start in a "better" region of the feature space. Data augmentation can help optimisation by reducing the risk overfitting to training samples. It can also learn a model invariant to information not useful for the task (e.g., rotation).

Motivated by these observations, we present an extensive ablation study of different preprocessing and data augmentation methods commonly found in the literature but whose usage is rarely empirically justified. Our goal is to establish the most appropriate methodological steps for segmentation of TBI lesions in CT. We explore the effects of spatial normalisation, intensity windowing, and skull-stripping on the final segmentation result. We also explore the effects of spatial normalisation vs. spatial data augmentation. We demonstrate that using intensity windowing can replace skull-stripping, and spatial data augmentation can replace spatial normalisation. We show the difference these methodological choices can make is up to 18% in the Dice similarity coefficient (DSC). To the best of our knowledge, this work is the first to report results for multi-class segmentation of contusion core, blood and oedema in TBI head CT.

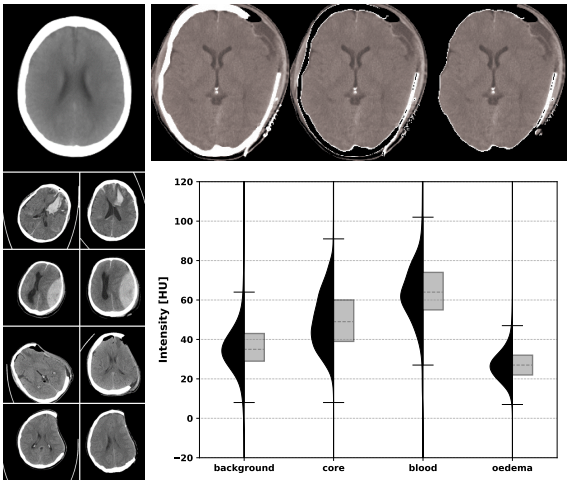

Fig. 1: **Left (a):** Effect of affine spatial normalisation to atlas space, CT atlas (top), before and after normalisation (left and right). **Top right (b):** Skull-stripping methods, from left to right: no skull-stripping; thresholding out the skull; level-set method. **Bottom right (c)**: Intensity distribution of classes inside the brain mask.

2 Methods

Dataset: We use 98 scans from 27 patients with moderate to severe TBI. We split the data into training and test (64/34), ensuring images from the same patient are in the same set. All scans have been manually annotated and reviewed by a team of experts to provide reference segmentations. We consider four classes: background; core; blood; and oedema. The core class includes contusion cores and petechial haemorrhages. The blood class includes subdural and extradural haematomas as well as subarachnoid and intraventricular haemorrhages.

Spatial normalisation: Since CNNs are not scale invariant, it is standard practice to resample all images to have the same physical (isotropic) resolution (e.g., $1\times1\times1$ mm). Given that brain CT is often highly anisotropic with high in-plane and low out-of-plane resolution, we may opt to resample to anisotropic resolution (e.g., $1\times1\times4$ mm) without loss of information while saving memory in the CNN's activations. Another preprocessing option is to perform spatial normalisation via registration to a reference frame, e.g., an atlas. Using affine transformations, we can remove inter-subject variability in terms of rotation and scaling which may be beneficial for the CNN. We investigate the effect of resampling and registration using three different settings: **1)** isotropic resolution of 1 mm; **2)** anisotropic resolution of $1\times1\times4$ mm; **3)** affine-to-atlas registration with 1 mm isotropic resolution. The atlas has been constructed from 20 normal CT scans that show no disease using an iterative unbiased atlas construction scheme [17], and subsequent alignment to an MNI MRI atlas (Fig. 1a).

Skull-stripping and intensity windowing: Skull-striping is commonly used in brain image analysis to eliminate unnecessary information by removing the skull and homogenising the background. Unlike MRI, CT intensities are roughly calibrated in HUs and have a direct physical interpretation related to the absorption of X-rays. A specific intensity value reflects the same tissue density regardless of the scanner. Air is defined as -1000 HUs, distilled water as 0, soft

tissue ranges between -100 and 300, while bone has larger values than soft tissue. Consider a lower and an upper bound that define the range of soft tissue. We test three different skull-stripping methods: **1)** no skull-stripping: we set intensities below the lower bound and above the upper bound to the lower and upper bound respectively; **2)** thresholding out the skull: we set values below the lower bound and above the upper bound to the lower bound; **3)** a level-set method (geodesic active contours [18]) to remove the skull followed by thresholding. Fig. 1b shows the effect of the three skull-stripping methods. We performed a visual check on all images to make sure the level-set method is not removing parts of the brain. We test two different intensity windows for the bounds and normalisation: **1)** a larger window range $[-100, 300]$; **2)** a smaller window $[-15, 100]$. Fig. 1c shows that intensity values of soft tissue fall well inside these windows. After skull-stripping and windowing, we normalise the intensity range to $[-1, 1]$. As seen in Fig. 1b, the brain-mask it not perfect in cases with a craniectomy, yet, this is a realistic scenario when we calculate automatic brain-masks for large datasets.

Data augmentation: We test the following settings: no augmentation; flipping the x axis; flipping the x and y axes; flipping the x and y axes combined with fixed rotations (multiples of 90°) of the same axes; flipping and fixed rotations of all axes; random affine transformations (scaling $\pm 10\%$; rotating xy randomly between $\pm 45°$; rotating xz and yz randomly between $\pm 30°$) combined with flipping the x axis; random affine transformations combined with flipping the x and y axes. We test these settings for both isotropic and affine spatial normalisation to study the effects of spatial augmentation vs. spatial normalisation.

Model architecture: Our main goal is to study the effects of preprocessing and hence we use the same architecture for all experiments[1]. We employ DeepMedic [5], a 3D CNN, with 3 parallel pathways that process an image at full resolution, three and five times downsampled. We use the residual version [19], but using 20 less feature maps per convolution layer. To see which results are model specific we re-run a subset of experiments with a 3D U-Net [20].

3 Results and Discussion

To assess which type of preprocessing is most effective we test all combinations of the aforementioned alternatives for spatial normalisation, intensity windowing and skull-stripping. For evaluation, we use the DSC calculated after transforming the prediction back into the original native image space (where the expert segmentation is defined). Thus, we guarantee an accurate comparison with the expert's segmentation. Fig. 3 shows the DSC of the foreground class for all preprocessing pipelines (flipping the x axis for augmentation). The foreground class consists of all lesion classes merged into one (after training) for an overall evaluation and comparison. Fig. 2 presents a visual comparison between manual segmentation and the prediction made by the best performing model. We can

[1] For the experiments with 1×1×4 resolution, we turn some of the isotropic kernels into anisotropic 3×3×1 kernels in order to obtain approximately the same receptive field as in the experiments with isotropic resolution. This did not affect the performance.

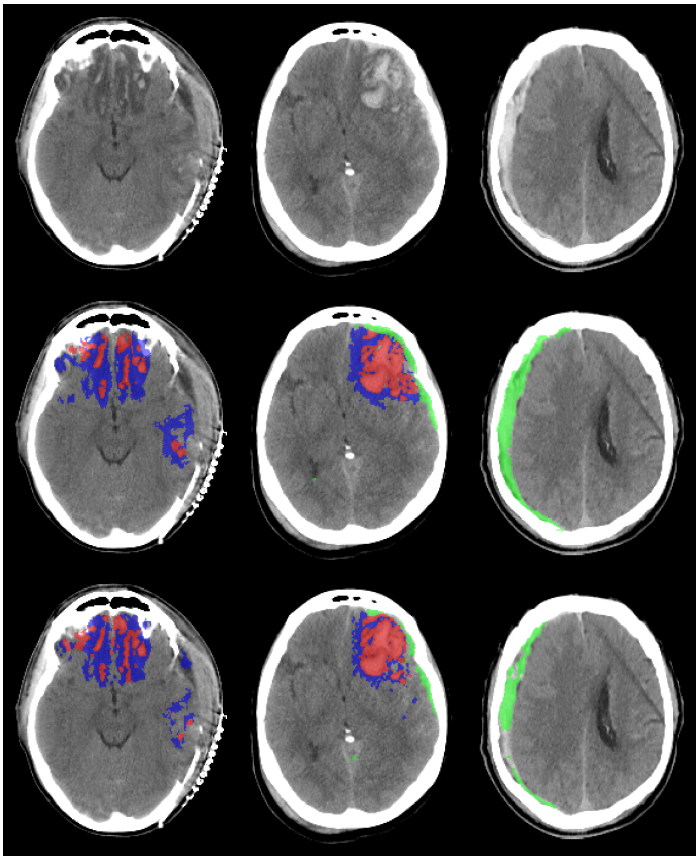

Fig. 2: Visual comparison for three cases. Top to bottom: image, manual and predicted segmentation. Red is contusion core, green is blood, and blue is oedema.

see that the TBI lesions have large inter-subject variability and intra-subject complexity, making for a difficult segmentation problem. Fig. 4 shows the result of paired Wilcoxon signed-rank tests to determine which performance differences are statistically significant ($p < 0.05$). Fig. 5a presents the per class DSC. Fig. 5b presents a subset of experiments replicated with a different model to determine if the results are model specific. Fig. 6 shows the results of spatial data augmentation for comparison with spatial normalisation.

Resampling images to isotropic resolution significantly improves performance. From Figs. 3 and 4 (purple) we observe that using anisotropic resolution of $1\times1\times4$ is consistently worse than using the other two spatial normalisation methods which use isotropic resolution. This contradicts the intuition that oversampling the out-of-plane axis is not necessary, even though it does not add information. CNNs have an architectural bias towards isotropic data which

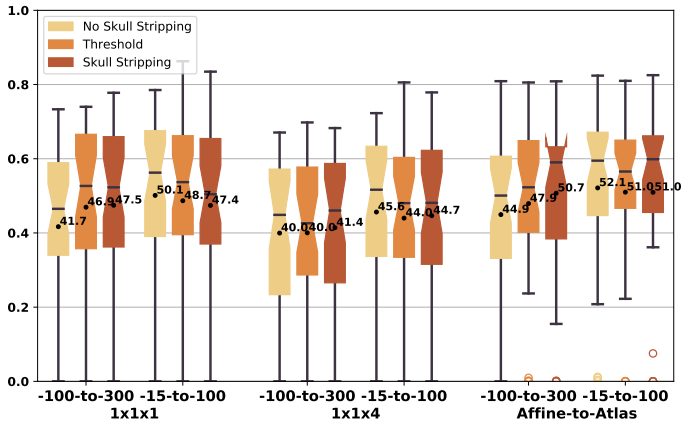

Fig. 3: Foreground DSC box plots for all preprocessing pipelines.

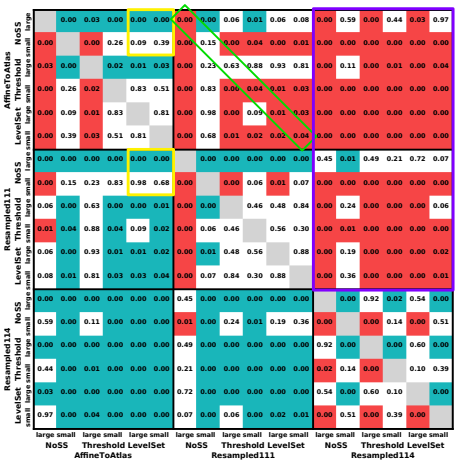

Fig. 4: Results of paired difference tests comparing prepossessing methods. For a p-value of 5% the y-label is statistically significantly better (red) or worse (blue) than the x-label. Yellow boxes: level-set skull stripping only helps performance when the large intensity window is used. Green box: controlling for other preprocessing steps, affine-to-atlas can provide small benefits. Purple box: anisotropic resolution consistently under-performs when compared to isotropic resolution.

is likely the cause for this result. The kernels are stacked such that it is expected the same amount of information to be present in each physical direction.

Skull-stripping can be replaced by windowing. From Fig. 4 (yellow) we can see that skull-stripping significantly helps performance when combined with a large intensity window. However, when combined with a small intensity window, skull-stripping does not offer a benefit over no skull-stripping. We observe the same for the second model (Fig. 5b) where the difference is also not statistically significant ($p = 0.2$). This indicates that what is important for the CNN is to remove intensity ranges that are not of interest (e.g., hyper-intense skull, either via windowing or skull stripping), allowing it to focus on the subtle intensity differences between lesions and healthy tissue. Therefore, skull-stripping may be replaced by an intensity window that limits extreme intensity values. Level-set based skull-stripping is susceptible to failure. Although we performed

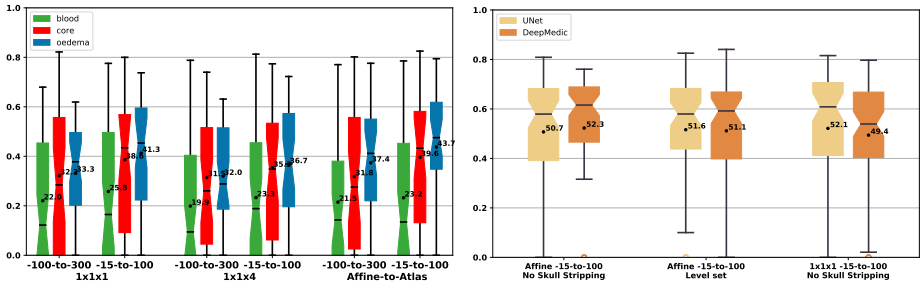

Fig. 5: **Left (a):** Per class DSC box plots preprocessing pipelines with no skull-stripping. **Right (b):** Foreground DSC box plots for two different models.

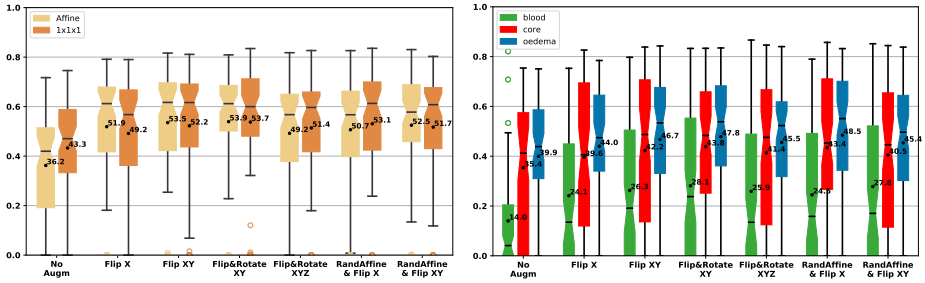

Fig. 6: **Left (a):** Foreground DSC box plots for all data augmentation methods. **Right (b):** Per class DSC box plots for all data augmentation methods.

visual checks to ensure quality, these are unfeasible on large scale settings such as the deployment of an automated segmentation pipeline. Conversely, windowing is more robust, hence it should be used instead.

Affine-to-atlas registration can be replaced by data-augmentation. Although Figs. 3, 4 (green) could lead us to believe that affine registration provides a small benefit over simply using isotropic resolution, when we look at Fig. 6 we see this is not the case. When we add more spatial augmentation to the two spatial normalisation methods (besides only flipping the x axis) their performance becomes comparable. Moreover, we see that this small benefit may not translate to different models (Fig. 5b). We conclude that we can make the network more robust to spatial heterogeneity with data augmentation instead of homogenising the input data. Like skull-stripping, affine-to-atlas registration can fail unpredictably during deployment of automatic pipelines. In contrast, spatial augmentation applied only during model training, and thus it does not constitute a possible point of failure after deployment. As a result, we recommend spatial data augmentation to be used instead of affine-to-atlas registration. Regardless, affine-to-atlas registration can still be useful for downstream analysis tasks such as studying the location of lesions across a patient cohort.

Additional findings. From Fig. 6 we observe that even though random affine augmentation serves its purpose, it did not perform better than fixed rotations. These transformations incur on a computational cost due to interpolation, and hence fixed rotations may be a better choice. We also observe that too much augmentation can start to hinder performance since flipping and rotating all axes performs worse than doing the same on just x and y.

We achieve a maximum DSC of 53.9±23.0 % (foreground). We can see from Figs. 5a and 6b that there is a large discrepancy between the performance of each class. The blood class has the worst performance likely due to the presence of hard to segment lesions such as subarachnoid haemorrhages. Surprisingly, the model performs best for oedema, one of the hardest lesion types to detect visually. Our results are not directly comparable with ones reported in the literature [10,11,12]. We use a different dataset, perform multi-class segmentation, and our labels include hard to segment lesions such as petechial and subarachnoid haemorrhages. Although the DSC obtained is not as high as in other similar applications (e.g. brain tumour segmentation on MRI), this goes to show the challenging nature of the problem and need to focus more effort on difficult tasks. Importantly, this is the first work reporting fine-grained multi-class segmentation of contusion core, blood and oedema in CT for patients with TBI.

4 Conclusion

We present an in-depth ablation study of common data preprocessing and augmentation methods for CT and show these methodological choices are key for achieving better segmentation performance with CNNs, with a difference of 18% DSC between the worst and best settings. Based on our results we make the following recommendations: **1)** using isotropic resolution is key **2)** choosing the correct intensity window for context and normalisation is superior to skull-stripping since it is simpler and more robust; **3)** affine-to-atlas registration can give small improvements, however, spatial data augmentation can achieve the same benefits while being more robust.

We hope our study will serve as a useful guide and help the community to make further progress on the clinically important task of head CT lesion segmentation. While our results on fine-grained TBI lesion segmentation are promising, we believe this study also shows that this task remains an open challenge and new approaches may be required to tackle this difficult problem. In the future, we aim to apply our findings to a larger dataset and to further fine-grain the segmented classes by separating SDH, EDH and SAH into separate classes.

Acknowledgments

This work is partially funded by a European Union FP7 grant (CENTER-TBI; Agreement No: 60215) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Agreement No: 757173,

project MIRA, ERC-2017-STG). EF is supported by the AXA Research Fund and UNL (CAID-50220140100084LI). KK is supported by the Presidents PhD Scholarship of Imperial College London. VFJN is supported by a Health Foundation/Academy of Medical Sciences Clinician Scientist Fellowship. DKM is supported by funding from the National Institute for Health Research (NIHR) through a Senior Investigator award and the Cambridge Biomedical Research Centre at the Cambridge University Hospitals National Health Service (NHS) Foundation Trust. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

References

1. D. K. Menon, K. Schwab, D. W. Wright, and A. I. Maas, "Position statement: Definition of traumatic brain injury," 2010.
2. A. I. R. Maas, D. K. Menon, and P. D. Adelson et al., "Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research," *The Lancet Neurology*, 2017.
3. J. P. Coles, "Imaging after brain injury," *BJA*, 2007.
4. A. Rao and C. Ledig, "Contusion segmentation from subjects with Traumatic Brain Injury: A random forest framework," *ISBI*, 2014.
5. 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," *MIA*, 2017.
6. T. Chan, "Computer aided detection of small acute intracranial hemorrhage on computer tomography of brain," *CMIG*, 2007.
7. M. R. Arbabshirani, B. K. Fornwalt, G. J. Mongelluzzo, J. D. Suever, B. D. Geise, A. A. Patel, and G. J. Moore, "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration," *npj Digital Medicine*, 2018.
8. J. R. Koikkalainen, J. M. P. Lötjönen, C. Ledig, D. Rueckert, O. S. Tenovuo, and D. K. Menon, "Automatic quantification of CT images for traumatic brain injury," *ISBI*, 2014.
9. S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study," *The Lancet*, 2018.
10. C.-C. Liao, F. Xiao, J.-M. Wong, and I.-J. Chiang, "Computer-aided diagnosis of intracranial hematoma with brain deformation on computed tomography," *CMIG*, 2010.
11. K. N. B. Prakash, S. Zhou, T. C. Morgan, D. F. Hanley, and W. L. Nowinski, "Segmentation and quantification of intra-ventricular/cerebral hemorrhage in CT scans by modified distance regularized level set evolution technique," *IJCARS*, 2012.
12. S. Jain, T. V. Vyvere, V. Terzopoulos, D. M. Sima, E. Roura, A. Maas, G. Wilms, and J. Verheyden, "Automatic quantification of computed tomography features in acute traumatic brain injury," *Journal of Neurotrauma*, 2019.
13. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, 1989.

14. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *MIA*, 2017.
15. K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, 1991.
16. J. Jo and Y. Bengio, "Measuring the tendency of CNNs to Learn Surface Statistical Regularities," *arXiv preprint*, 2017.
17. S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage*, 2004.
18. V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *IJCV*, 1997.
19. K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. V. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "DeepMedic for brain tumor segmentation," in *LNSC*, 2016.
20. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *LNSC*, 2015.