Original Article

Assessment of Optimizers and their Performance in Autosegmenting Lung Tumors

Prabhakar Ramachandran, Tamma Eswarlal[1], Margot Lehman, Zachery Colbert

Department of Radiation Oncology, Cancer Services, Princess Alexandra Hospital, Woolloongabba, Queensland, Australia, [1]Department of Engineering Mathematics, College of Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

Abstract

Purpose: Optimizers are widely utilized across various domains to enhance desired outcomes by either maximizing or minimizing objective functions. In the context of deep learning, they help to minimize the loss function and improve model's performance. This study aims to evaluate the accuracy of different optimizers employed for autosegmentation of non-small cell lung cancer (NSCLC) target volumes on thoracic computed tomography images utilized in oncology. **Materials and Methods:** The study utilized 112 patients, comprising 92 patients from "The Cancer Imaging Archive" (TCIA) and 20 of our local clinical patients, to evaluate the efficacy of various optimizers. The gross tumor volume was selected as the foreground mask for training and testing the models. Of the 92 TCIA patients, 57 were used for training and validation, and the remaining 35 for testing using nnU-Net. The performance of the final model was further evaluated on the 20 local clinical patient datasets. Six different optimizers, namely AdaDelta, AdaGrad, Adam, NAdam, RMSprop, and stochastic gradient descent (SGD), were investigated. To assess the agreement between the predicted volume and the ground truth, several metrics including Dice similarity coefficient (DSC), Jaccard index, sensitivity, precision, Hausdorff distance (HD), 95th percentile Hausdorff distance (HD95), and average symmetric surface distance (ASSD) were utilized. **Results:** The DSC values for AdaDelta, AdaGrad, Adam, NAdam, RMSprop, and SGD were 0.75, 0.84, 0.85, 0.84, 0.83, and 0.81, respectively, for the TCIA test data. However, when the model trained on TCIA datasets was applied to the clinical datasets, the DSC, HD, HD95, and ASSD metrics showed a statistically significant decrease in performance compared to the TCIA test datasets, indicating the presence of image and/or mask heterogeneity between the data sources. **Conclusion:** The choice of optimizer in deep learning is a critical factor that can significantly impact the performance of autosegmentation models. However, it is worth noting that the behavior of optimizers may vary when applied to new clinical datasets, which can lead to changes in models' performance. Therefore, selecting the appropriate optimizer for a specific task is essential to ensure optimal performance and generalizability of the model to different datasets.

Keywords: Autosegmentation, computed tomography, deep learning, optimization algorithm

Received on: 25-04-2023 **Review completed on: 06-05-2023** **Accepted on: 14-05-2023** **Published on: 29-06-2023**

Introduction

According to the World Health Organization, lung cancer is one of the most common types of cancer with the highest mortality rate. In general, nearly 80%–85% of all lung cancers present as non-small cell lung cancer (NSCLC), with the remaining being small cell lung cancers (SCLC).[1,2] SCLC is the more aggressive of the two and is more likely to migrate to other parts of the body. Treatments for lung cancer may include surgery, radiation therapy, and/or chemotherapy. Prior to radiation therapy for lung cancer, several steps are required, including imaging, treatment planning, quality assurance, and treatment verification. The tumor lesion is delineated prior to treatment planning to ensure that the intended dose conforms to the target volume with minimal exposure to surrounding healthy tissue. Delineation of organs at risk (OARs) and tumor volumes is a time-consuming process, in which the quality and accuracy of contours depend on the skill set of the observer.[3] Artificial intelligence (AI)-based autosegmentation models have been developed to overcome this limitation and to improve the consistency and efficiency of lung cancer treatment planning. AI-based autosegmentation models vastly

Access this article online

Quick Response Code:

Website: www.jmp.org.in

DOI: 10.4103/jmp.jmp_54_23

Address for correspondence: Dr. Prabhakar Ramachandran, Department of Radiation Oncology, Cancer Services, Princess Alexandra Hospital, Woolloongabba 4102, Queensland, Australia.
E-mail: prabhakar.ramachandran@health.qld.gov.au

How to cite this article: Ramachandran P, Eswarlal T, Lehman M, Colbert Z. Assessment of optimizers and their performance in autosegmenting lung tumors. J Med Phys 2023;48:129-35.

improve the reproducibility of delineations while reducing the time spent performing manual segmentation.[4]

Deep learning techniques have become increasingly popular in the medical field for developing algorithms that automatically detect abnormalities, classify medical images, segment organs, and detect abnormal masses.[5] Manual segmentation from medical images can be challenging for several reasons. The complex and irregular shape of lungs and their varying densities and textures make it difficult to accurately delineate the tumor boundary. Moreover, lung tumors are heterogeneous, and can have different characteristics within the same volume, which makes it challenging to distinguish the tumor from adjacent healthy tissues. The quality of medical images is also affected by acquisition techniques and parameters, resulting in low-resolution images with minimal contrast and artifacts. Furthermore, lung tumors are prone to movement due to breathing, and depending on their location in the thoracic cavity and breaths per minute, it can lead to artifacts on the scans and may also be subjected to deformation.

The accuracy of deep learning models is influenced by several parameters, including the optimizer, batch sizes, sample sizes, weights and biases, learning rates, and epochs. Optimizers are particularly important as they reduce losses and guide the model toward the global loss minimum. Multiple optimizers are available in deep learning, each with its own advantages and disadvantages. For example, while stochastic gradient descent (SGD) is a widely used optimizer, it may converge slowly for certain datasets.[6] SGD utilizes the gradient of the loss function to update the model's parameters and is suitable for both small and large datasets. PyTorch offers several extensions to SGD, including momentum-based SGD using Nesterov momentum, with weight decay to improve its performance on certain tasks.[7]

Another optimizer, RMSprop, is similar to SGD but with some adjustments that make it more likely to be more effective.[8] However, like other optimization methods, RMSprop can be unstable and may struggle to converge to the global minima. Adam, a relatively new optimizer, combines the strengths of both SGD and RMSprop and is generally faster than SGD while being less sensitive to hyperparameters.[9] Adam is more stable than other optimization methods and performs well on a wide range of tasks. It is particularly useful for training models with large datasets and high-dimensional parameter spaces. Despite the overall effectiveness of the Adam optimizer, it is still susceptible to the potential of not converging to the global minimum of the optimization problem. In addition to these optimizers, AdaGrad, AdaDelta, and NAdam have also been shown to be effective in deep learning.[10-13]

The nnU-Net is a deep learning architecture that automatically configures itself for medical image segmentation. It is a 3D U-Net architecture extensively used in biomedical image segmentation.[14,15] It includes several enhancements over simpler 3D U-Net models, such as optimized preprocessing, training strategies, and architecture adaptations. nnU-Net has demonstrated its potential for volume autosegmentation on medical images by winning several segmentation challenges. To handle noise in medical imaging data and achieve faster convergence during training, nnU-Net uses SGD with Nesterov momentum of 0.9 as the default optimizer. Furthermore, using Nesterov momentum reduces the oscillations typically associated with momentum-based optimization methods. nnU-Net divides hyperparameters into fixed, rule-based, and empirical configurations.

The location, heterogeneity, and presence of surrounding edema pose challenges for developing a robust autosegmentation model for lung tumors. This study focused on using multiple optimizers, including SGD with Nesterov momentum, Adam, AdaDelta, AdaGrad, NAdam, and RMSprop, for autosegmentation of lung tumors belonging to a publicly available "The Cancer Imaging Archive" (TCIA) datasets, and applying the trained model to local clinical lung computed tomography (CT) datasets.

Materials and Methods

One hundred and twelve lung cancer CT patient datasets, consisting of 92 TCIA patients[16] and 20 local clinical lung CT datasets, were included in this study. The gross tumor volume without any additional margin was considered the lesion for training and testing the models. Of the 92 TCIA patients, 57 were used for training and validation, and the remaining 35 cases were reserved for testing. The final model was additionally tested on the 20 local clinical datasets. The purpose was to evaluate the performance of a model trained on a publicly available dataset and its relative behavior when applied to local clinical cases. The training and testing of the autosegmentation models were performed on a dedicated Linux Ubuntu server housing a GeForce RTX 3090 Ti 24 GB graphics processing unit with 96 GB RAM.

The nnU-Net architecture was used to generate the lung autosegmentation models. The CT DICOM image and RT structure set files were converted to NifTi format. The images and masks were mapped together with the assistance of a json file. The learning rates for SGD with Nesterov momentum, Adam, AdaGrad, AdaDelta, NAdam, and RMSprop were 10^{-2}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-5}, and 10^{-3}, respectively. The training was performed using the nnU-Net 3D full-resolution architecture, which utilizes images at their full resolution. The average lung lesion volumes in the training and test datasets from TCIA, consisting of 57 and 35 patients, respectively, were 66.86 ± 80.01 cc (ranging from 2.85 to 400.82) and 69.59 ± 123.26 cc (ranging from 2.65 to 661.35). The average volume of local clinical lung lesions included in the study was 47.38 ± 76.17 cc (ranging from 1.3 to 284.34 cc). To determine if the trained model could predict lesions outside the trained minimum volume, the study included three clinical lung cases with volumes below the minimum trained volume (1.32 cc, 1.64 cc, and 2.35 cc). A *t*-test for two independent means with a significance level of $P < 0.05$ was used to evaluate

the statistical difference between the TCIA test and clinical datasets. To assess the agreement between the ground truth and predicted volume, the Dice similarity coefficient (DSC), Jaccard index, precision, Hausdorff distance (HD), 95th percentile Hausdorff distance (HD95), and average symmetric surface distance (ASSD) were calculated. To compute the evaluation metrics and volume of the lesions, a Python script was created that employed parallel processing to expedite the computation process. Prior to calculating the metrics for assessing the agreement between the ground truth and predicted lesions, both volumes were converted into binary objects.

DSC is a statistical metric used to evaluate the similarity or overlap between two sets of data or objects.[17] In the context of image segmentation, it is commonly used to quantify the similarity between a ground truth image (A) and a segmented/predicted image (B). It is calculated as the ratio of twice the intersection of A and B to the total sum of A and B [Figure 1]. The range of values for DSC is between 0 and 1, with a value of 1 indicating a perfect match between the two sets.

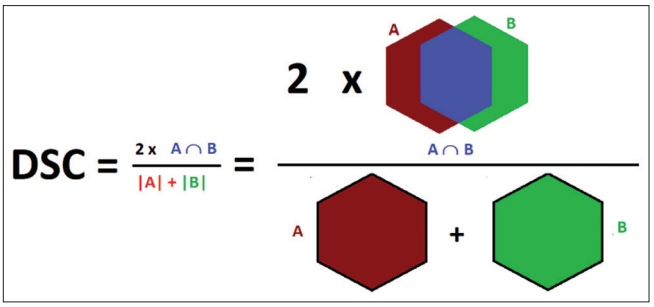

Figure 1: Dice similarity coefficient of two objects A and B

Like DSC, the Jaccard index, also known as the intersection over union, is a statistical metric utilized to assess the similarity or overlap between two sets of data. It is computed by dividing the intersection of sets A and B by the union of sets A and B, as shown in Figure 2.

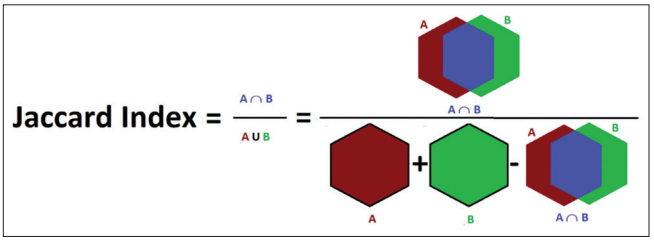

Figure 2: Jaccard index of two objects A and B

HD is a mathematical concept used to measure the agreement between two sets of points in a metric space.[18] Specifically, it measures the greatest distance between any point in one set to the closest point in the other set. In other words, the HD between two sets of points is the maximum distance of a point in one set to its nearest point in the other set. For two nonempty subsets A and B of a metric space M, the directed HD h (A, B) from set A to set B is defined as:

$$h(A,B) = \max_{a \in A} \min_{b \in B} \| a - b \|$$

Figure 3 illustrates the ground truth (A) and predicted (B) volumes' boundary along with the minimum distances from boundary points/pixels in A to B, and vice versa. The HD between sets/objects A and B is then defined as the maximum of the two directed HDs, h (A, B) and h (B, A):

HD (A, B) = max (h(A, B), h(B, A))

The metric HD (A, B) captures the maximum dissimilarity or separation between the two objects/volumes A and B. The HD95 is a metric used to represent the 95th percentile of the distances between the surface points of the ground truth object (A) and the corresponding points on the predicted object (B). In other words, it measures the maximum distance between the two objects such that 95% of the surface points of the ground truth object are within this distance from the corresponding surface points of the predicted object. HD95 is a robust metric that is less sensitive to small outliers than the traditional HD metric, as it excludes the largest 5% of values.

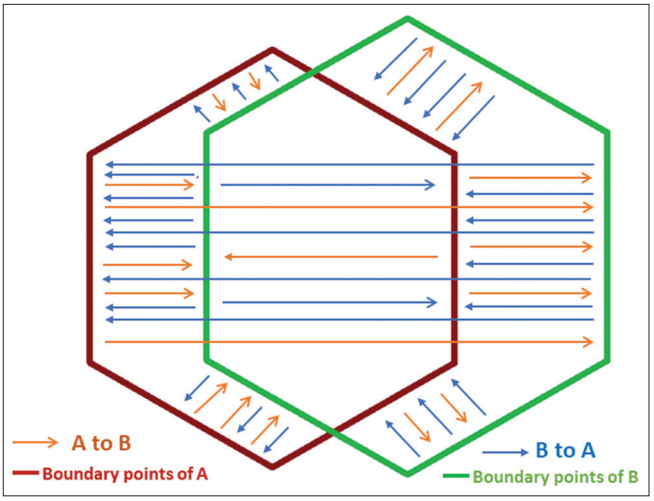

Figure 3: Illustration of the methodology for computing HD and ASSD. HD: Hausdorff distance, ASSD: Average symmetric surface distance

The higher the HD95 value, the greater the dissimilarity or deviation between the two objects A and B.

The average surface distance (ASD) is a metric used to measure the average distance between points in objects or regions of interest, such as A and B, while minimizing the sum of distances. To calculate ASD, the mean surface distance of object A and object B is computed separately in both directions. Subsequently, the ASSD is determined by taking the mean of the ASDs in both directions.[19] This means that the distance is calculated both from the ground truth volume surface to the predicted volume surface and vice versa, and then averaged over all boundary points. ASSD is generally more robust to outliers compared to the HD alone. It considers the average distance between the surfaces of the two objects, which tends to reduce the impact

of isolated extreme distances or outliers that may exist when comparing two volumes.

ASSD (A, B) = (ASD [A to B] + ASD [B to A])/2

Results

The lung tumor prediction model, trained on a dataset of 57 patients, demonstrated accurate tumor prediction for all 35 test patients from TCIA using AdaDelta, AdaGrad, Adam, NAdam, RMSprop, and SGD optimization algorithms. Figure 4 shows the predicted and ground truth masks from the test TCIA thoracic datasets which had the highest DSC value for most models.

Table 1 presents the mean DSC, Jaccard index, precision, HD, HD95, and ASSD results for the trained models. Except for AdaDelta, all models exhibited average DSC values > 0.80. The Jaccard index followed a similar pattern to the DSC. All models displayed high precision, with values above 0.86. Notably, AdaDelta and SGD showed high HD, HD95, and ASSD values, indicating significant shape differences between the ground truth and predicted volumes.

For the clinical datasets, most of the models trained on the dataset of 57 patients were unable to predict the lesions in the three cases (with ground truth volumes of 1.32, 1.64, and 2.35 cc) that had volumes less than the minimum trained lesion volume (2.85 cc). As a result, these three volumes were not taken into consideration for the analysis. Figure 5 shows the predicted and ground truth masks from the clinical lung thoracic datasets with a high DSC value.

Table 2 presents the mean DSC, Jaccard index, Precision, HD, HD95, and ASSD for the various optimizers' models. Consistent with the DSC, the Jaccard index also decreased for all optimizers when the TCIA-trained model was applied to the clinical datasets. All models exhibited a significant variation in HD, HD95, and ASSD when the trained model was used on the clinical datasets, indicating data heterogeneity between the TCIA and the clinical datasets.

Figure 6 illustrates the trend of the average DSC, indicating that the models trained on the TCIA datasets performed well on those datasets; however, there was a decrease in the DSC values when applied to the local clinical dataset. The *t*-test for two independent means revealed significant differences ($P < 0.05$) between the DSC calculated on the TCIA test data and the clinical lung datasets.

The results also demonstrate that RMSprop and SGD performed relatively better on the local clinical datasets, while all other optimizers exhibited a significant decrease in DSC when the model was transferred to the clinical datasets.

Discussion

Deep learning models have shown significant potential in automating the segmentation of lung tumors and OARs in (CT) images.[20-22] This can improve the accuracy and efficiency

Table 1: Mean performance metrics comparison of optimizers on 35 The Cancer Imaging Archive test datasets

	AdaDelta	AdaGrad	Adam	NAdam	RMSprop	SGD
DSC	0.75	0.84	0.85	0.84	0.83	0.81
Jaccard index	0.63	0.74	0.75	0.73	0.73	0.72
Precision	0.87	0.88	0.89	0.88	0.86	0.89
HD (mm)	12.79	9.95	9.32	9.12	9.58	14.18
HD95 (mm)	7.45	4.96	4.38	4.92	4.52	8.79
ASSD (mm)	2.15	1.41	1.31	1.43	1.23	2.89

HD: Hausdorff distance, HD95: 95th percentile HD, ASSD: Average symmetric surface distance, SGD: Stochastic gradient descent, DSC: Dice similarity coefficient, Adam: Adaptive Moment Estimation, NAdam: Nesterov-Accelerated Adaptive Moment Estimation, RMSprop: Root Mean Square Propagation

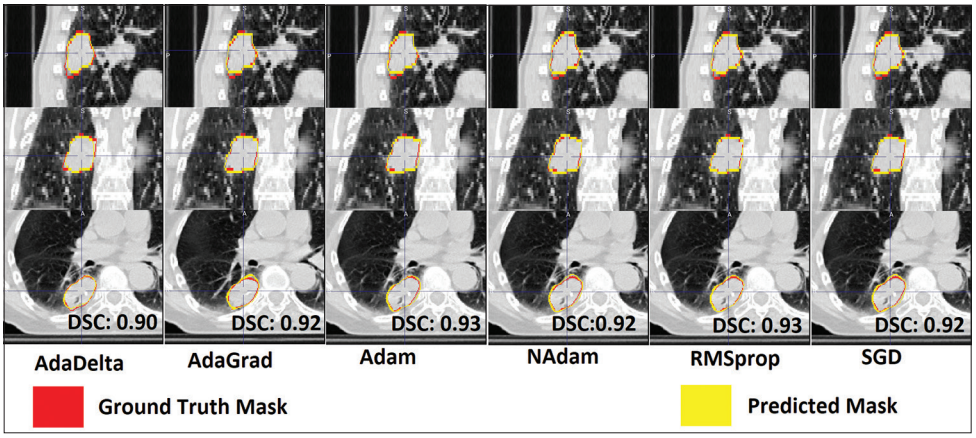

Figure 4: Ground truth and predicted TCIA Test data volume comparison for all six optimizers (sagittal, coronal and axial views). TCIA: The Cancer Imaging Archive

Table 2: Mean performance metrics comparison of optimizers on 17 clinical cases

	AdaDelta	AdaGrad	Adam	NAdam	RMSprop	SGD
DSC	0.67	0.72	0.77	0.69	0.80	0.76
Jaccard index	0.53	0.60	0.64	0.56	0.67	0.64
Precision	0.87	0.92	0.91	0.83	0.90	0.93
HD (mm)	14.83	15.78	12.75	14.06	11.77	12.25
HD95 (mm)	9.36	9.50	7.08	7.52	5.92	7.08
ASSD (mm)	2.94	2.94	2.18	2.29	2.05	2.24

HD: Hausdorff distance, HD95: 95th percentile HD, ASSD: Average symmetric surface distance, SGD: Stochastic gradient descent, DSC: Dice similarity coefficient, Adam: Adaptive Moment Estimation, NAdam: Nesterov-Accelerated Adaptive Moment Estimation, RMSprop: Root Mean Square Propagation

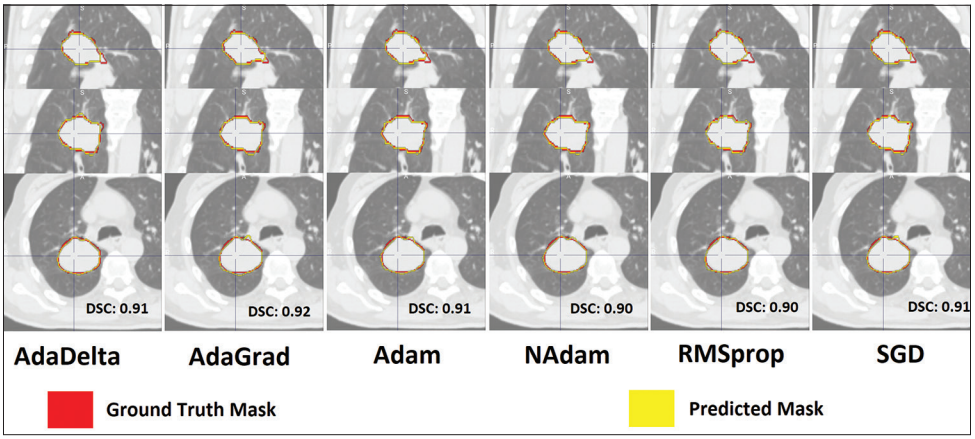

Figure 5: Ground truth and predicted clinical data volume comparison for all six optimizers (sagittal, coronal and axial views)

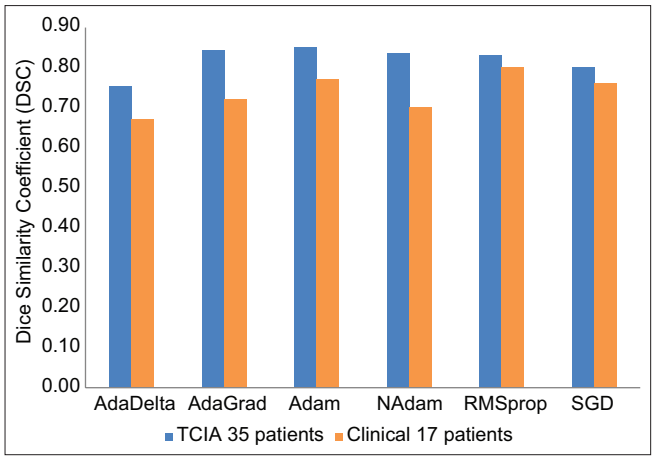

Figure 6: Mean DSC comparison between the TCIA test dataset and clinical dataset. DSC: Dice similarity coefficient, TCIA: The Cancer Imaging Archive

of treatment planning, particularly in radiation therapy. Autosegmentation of lung tumors is challenging due to factors such as location, variability in shape and size, interobserver variability,[23,24] tumors surrounded by atelectasis,[25] low contrast with surrounding tissues, volume changes introduced by respiration, and image artifacts.

Our study primarily focused on the performance of different optimizers for autosegmenting NSCLC tumors. It demonstrated that most optimizers produced decent DSC values above 0.80, with the exception of AdaDelta. When applying the TCIA-trained model to the clinical datasets, we observed a significant drop in DSC values for most optimizers. RMSprop and SGD showed a minimal decline in DSC values. An increase in the mean HD indicates greater, indicating differences between the ground truth and predicted volumes. This suggests that the model performance decreased compared to the predictions on TCIA datasets. When comparing the mean HD and mean HD95 results between the TCIA and clinical datasets, TCIA HD and HD95 values were lower than the clinical datasets, except for SGD. Furthermore, HD is skewed by outliers whereas HD95 indicates the distance below which 95% of the points fall, effectively capturing the majority of the differences between the ground truth and predicted volumes. This observation suggests that the HD95 metric provides a more robust measure of distance, as it is less affected by outliers and extreme values. On average, there was a 33% increase in the mean HD95 across all optimizers when evaluating the clinical datasets compared to the TCIA test datasets. Similarly, in line with the HD95 metric, the mean ASSD exhibited a 40% increase when comparing the clinical datasets to the TCIA datasets. In the present study, we observed that the Jaccard index consistently yielded lower values compared to the DSC when evaluating various optimizers with both the TCIA and clinical datasets. This discrepancy can be attributed to the fundamental differences in how these metrics calculate similarity. The Jaccard index determines similarity by dividing the size of the intersection of two sets by the size

of their union. Since the intersection is always a subset of the union, the Jaccard index tends to produce lower values compared to the DSC. On the other hand, the DSC places greater emphasis on the size of the intersection by considering twice the intersection's size divided by the sum of both set sizes. As a result, this formulation can generate higher values compared to the Jaccard index. Additionally, the Jaccard index is sensitive to the size of the sets being compared.[26] When the sets are small, there is a higher chance of a relatively larger intersection, leading to higher Jaccard index values. In our study, we had 92 TCIA datasets for training and testing, which helped ensure reliable comparisons. In contrast to the Jaccard index, the DSC is less sensitive to set size, resulting in more balanced values for different set sizes. It is important to note that the specific values obtained from these similarity metrics will depend on the characteristics of the sets being compared and the specific problem domain. Zhang *et al*. reported that their modified version of ResNet yielded a DSC of 0.73, Jaccard index of 0.68, and a true positive rate of 0.71 when trained on a dataset of 400 NSCLC patients.[27] Our study revealed true positive rates of 0.72, 0.83, 0.84, 0.83, 0.85, and 0.79 with AdaDelta, AdaGrad, Adam, NAdam, RMSprop, and SGD optimizers for the TCIA test datasets. In the clinical datasets, the best true positive rate (0.74) was observed with RMSprop. Bi *et al*.'s study using a deep dilated residual network on 418 CT data showed an average DSC of 0.75.[28] However, combining positron emission tomography and CT has been shown to improve the DSC to 0.85.[29] Lung autosegmentation was also studied on stereotactic ablative radiotherapy datasets, yielding a DSC of 0.71.[30] While these studies show promising results for autosegmentation models in contouring lung tumors, it is essential to note that they can still be subject to limitations, such as sensitivity to variations in image quality, tumor size, and tumor shape. Additionally, the performance of these models is highly dependent on the quality and diversity of the training data. As a result, it is generally recommended to use autosegmentation as a starting point, with manual adjustments and verification by experienced radiation oncologists to ensure accurate tumor delineation. The memory size of a GPU determines the maximum size of datasets and models that can be stored and processed simultaneously.[31] When training deep learning models on large datasets, the model parameters and intermediate activations consume a significant portion of the available memory. Several methods have been proposed to reduce memory consumption, which can assist in training large datasets.[32,33] In this study, it took approximately 20 hours to run 1000 epochs for one optimizer. nnU-Net dynamically adjusts the batch size, patch size, and number of pooling operations for each axis based on the available memory resources. It optimizes these parameters to balance computational efficiency and memory consumption.

It is essential to acknowledge that the choice of optimization algorithm for autosegmentation can depend on various factors, including the target volume's type, size, and location, the dataset size, and the neural network architecture. While the presented results suggest that Adam performed well on the TCIA dataset and RMSprop and SGD displayed minimal variation in DSC between TCIA and clinical datasets, it should be noted that these optimizers may perform well for this specific task, but their performance may differ for different type of tumors or OARs.

Conclusion

The choice of optimizer in deep learning is a critical factor that can significantly impact the performance of autosegmentation models. However, it is worth noting that the behavior of optimizers may vary when applied to clinical datasets, which can lead to differences in model performance. Therefore, selecting the appropriate optimizer for a specific task is essential to ensure optimal performance and generalizability of the model to different datasets.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Zarogoulidis P, Chatzaki E, Porpodis K, Domvri K, Hohenforst-Schmidt W, Goldberg EP, *et al*. Management of lung cancer patients. J Thorac Dis 2013;5 Suppl 4:S416-23.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023;73:17-48.
3. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. Radiother Oncol 2016;121:169-79.
4. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, *et al.* Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41:050902.
5. Ramachandran P, Seshadri V, Perrett B, Mehta A, Fontanarosa D, Pinkham M, *et al*. Role of artificial intelligence in automatic segmentation of brain metastases for radiotherapy. InArtificial Intelligence in Cancer Diagnosis and Prognosis, Volume 3: Brain and prostate cancer 2022 Oct 1. IOP Publishing; P. 4-1 to 4-23.
6. Bottou L. Stochastic gradient descent tricks. In: Montavon G, Orr GB, Müller, KR. (eds) Neural Networks: Tricks of the Trade. 2nd edition. Springer, Berlin, Heidelberg; 2012. p. 421-36.
7. Ketkar N, Moolayil J. Introduction to pytorch. In: Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch. Apress LP, New York; 2021. p. 27-91.
8. Xu D, Zhang S, Zhang H, Mandic DP. Convergence of the RMSProp deep learning method with penalty for nonconvex optimization. Neural Netw 2021;139:17-23.
9. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv 2014;1412.6980. p. 1-15.
10. Lydia A, Francis S. Adagrad-an optimizer for stochastic gradient descent. Int J Inf Comput Sci 2019;6:566-8.
11. Zeiler MD. Adadelta: An adaptive learning rate method. arXiv 2012;1212.5701.
12. Tato A, Nkambou R. Improving Adam Optimizer. Workshop track - ICLR 2018:1-4.
13. Ruder S. An overview of gradient descent optimization algorithms. arXiv 2016;1609.04747.
14. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203-11.
15. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International

Conference, Proceedings, Part III. Munich, Germany: Springer International Publishing; 2015. p. 234-41.
16. NSCLC-Radiomics – The Cancer Imaging Archive (TCIA) Public Access – Cancer Imaging Archive Wiki. Available from: https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics. [Last accessed on 2022 Jun 05].
17. Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. The Royal Danish Academy of Sciences 1948;5:1-34.
18. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Tran Pattern Analysis Mach intell 1993;15:850-63.
19. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. deep learning for brain MRI segmentation: State of the art and future directions. J Digit Imaging 2017;30:449-59.
20. Johnston N, De Rycke J, Lievens Y, van Eijkeren M, Aelterman J, Vandersmissen E, *et al.* Dose-volume-based evaluation of convolutional neural network-based auto-segmentation of thoracic organs at risk. Phys Imaging Radiat Oncol 2022;23:109-17.
21. Mehta A, Lehman M, Ramachandran P. Autosegmentation of lung computed tomography datasets using deep learning U-Net architecture. J Cancer Res Ther 2023;19:289-98.
22. Yang J, Veeraraghavan H, Armato SG 3rd, Farahani K, Kirby JS, Kalpathy-Kramer J, *et al.* Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Med Phys 2018;45:4568-81.
23. Persson GF, Nygaard DE, Hollensen C, Munck af Rosenschöld P, Mouritsen LS, Due AK, *et al.* Interobserver delineation variation in lung tumour stereotactic body radiotherapy. Br J Radiol 2012;85:e654-60.
24. Zhao B, Schwartz LH, Moskowitz CS, Ginsberg MS, Rizvi NA, Kris MG. Lung cancer: Computerized quantification of tumor response – Initial results. Radiology 2006;241:892-8.
25. Mercieca S, Belderbos JSA, van Herk M. Challenges in the target volume definition of lung cancer radiotherapy. Transl Lung Cancer Res 2021;10:1983-98.
26. Grady CL, Rieck JR, Nichol D, Rodrigue KM, Kennedy KM. Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. Hum Brain Mapp 2021;42:204-19.
27. Zhang F, Wang Q, Li H. Automatic segmentation of the gross target volume in non-small cell lung cancer using a modified version of ResNet. Technol Cancer Res Treat. 2020;19:1533033820947484. p. 1-9.
28. Bi N, Wang J, Zhang T, Chen X, Xia W, Miao J, *et al.* Deep learning improved clinical target volume contouring quality and efficiency for postoperative radiation therapy in non-small cell lung cancer. Front Oncol 2019;9:1192.
29. Zhao X, Li L, Lu W, Tan S. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. Phys Med Biol 2018;64:015011.
30. Wong J, Huang V, Giambattista JA, Teke T, Kolbeck C, Giambattista J, *et al.* Training and validation of deep learning-based auto-segmentation models for lung stereotactic ablative radiotherapy using retrospective radiotherapy planning contours. Front Oncol 2021;11:626499.
31. Chen T, Xu B, Zhang C, Guestrin C. Training deep nets with sublinear memory cost. arXiv 2016;1604.06174.
32. Meng C, Sun M, Yang J, Qiu M, Gu Y. Training Deeper Models by GPU Memory Optimization on TensorFlow. Proceedings of ML Systems Workshop in NIPS; Location: Long Beach, CA;2017.
33. Bognar J, Ramachandran P. Autosegmentation of brain metastases using 3D FCNN models and methods to manage GPU memory limitations. Biomed Phys Eng Express 2022;8:065027.