Automatic segmentation of vestibular schwannoma from MRI images using deep learning

Lappeenranta–Lahti University of Technology LUT

Bachelor’s Programme in software engineering, Bachelor's thesis

2023

Joonas Liedes

Examiner: Associate Professor, Antti Knutas, D.Sc.

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Science

Software Engineering

Joonas Liedes

Automatic segmentation of vestibular schwannoma from MRI images using deep learning

Bachelor's thesis

2023

23 pages, 4 figures, 1 table

Examiner: Associate Professor, Antti Knutas, D.Sc.

Keywords: deep learning, open-source, medical computer vision

Deep learning (DL) has been used successfully for medical image interpretation, traditionally a labour-intensive task that requires a high level of specialised training. A growing body of evidence suggests that the incorporation of DL systems into medical image analysis can provide great efficiency benefits and better patient outcomes. However, it's broadscale research and implementation can still be impeded by steep hardware requirements and absence of freely available quality datasets. This thesis aims to demonstrate the feasibility of using open-sourced data and computing resources for building DL systems with real-world impact by solving a challenging medical computer vision problem – the automatic segmentation of a benign tumour, vestibular schwannoma (VS), from MRIs.

With this aim, a 3D U-Net inspired DL model was built and trained on a publicly available dataset containing MRIs of 242 VS patients. The data was split into training and testing subsets with an 80/20-ratio. A 5-fold cross-validation with the training data was carried out to address the limited number of samples. Based on the cross-validation the model with the lowest standard deviation in terms of Dice score was selected to be trained further with the complete training data. This model was then evaluated on the test set achieving a mean Dice score of 0.85 ± 0.21. The model training and testing, excluding the cross-validation, was additionally carried out in Google Colab utilising the free distributed GPU resource without issues.

These results support previous findings where automatic segmentation of VS with DL has been found accurate, efficient and of clinical value. Moreover, leveraging open-source resources for building DL models with potential real-world impact was found viable.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT Teknis-luonnontieteellinen

Tietotekniikan koulutusohjelma

Joonas Liedes

Vestibulaari schwannooman automaattisegmentointi magneettikuvista syväoppimista hyödyntäen

Tietotekniikan kandidaatintyö

2023

23 sivua, 4 kuvaa, 1 taulukko

Tarkastaja: Apulaisprofessori, Antti Knutas, D.Sc.

Avainsanat: syväoppiminen, avoin lähdekoodi, lääketieteellinen konenäkö

Syväoppimista on käytetty menestyksekkäästi lääketieteellisten kuvien tulkinnassa viime aikoina. Tätä tehtävää on tyypillisesti pidetty työläänä ja haastavana, mikä vaatii tulkitsevalta lääkäriltä pitkän erikoistumisen. Tutkimusnäyttö syväoppimisen tuomasta tehokkuusedusta lääketieteellisessä kuva-analyysissä ja sen myötä parantuneista potilasennusteista on kasvussa, mutta sen integraatio sairaala-arkeen on yhä rajallista. Syväoppimisen tutkimusta ja laaja-alaisempaa käyttöönottoa rajoittavat osaltaan koulutusprosessin laitteistovaatimukset sekä riittävän laadukkaiden ja kattavien koulutusaineistojen saatavuus. Tämän tutkielman tavoitteena on osoittaa avoimesti saatavilla olevien resurssien käyttökelpoisuus syväoppimisjärjestelmien kehityksessä rakentamalla vestibulaari schwannoomaa automaattisesti magneettikuvista segmentoiva järjestelmä näitä hyödyntäen.

Tämän tavoitteen saavuttamiseksi rakennettiin 3D U-Net tyyppinen syväoppimismalli. Malli koulutettiin hyödyntäen avoimesti saatavilla olevaa aineistoa, joka sisälsi 242 vestibulaari schwannoomaa sairastavan potilaan magneettikuvat. Aineisto jaettiin koulutus- ja testausjoukkoihin 80/20-suhteessa. Rajallinen aineisto pyrittiin huomioimaan käyttämällä viisinkertaista ristiinvalidointia koulutusjoukolle. Tämän perusteella valittiin malli, jolla oli alhaisin keskihajonta Dice-pisteiden osalta. Tämä malli koulutettiin edelleen käyttäen koko koulutusjoukkoa ja arvioitiin sitten hyödyntäen testijoukkoa, jolle keskimääräinen Dice-pistemäärä oli 0.85 ± 0.21. Mallin koulutus ja testaus, pois lukien ristiinvalidointi, suoritettiin ongelmitta hyödyntäen Googlen Colab -palvelun ilmaista prosessointiresurssia.

Nämä tulokset tukevat aiempia löydöksiä, joissa syväoppiminen on osoittautunut tarkaksi ja tehokkaaksi menetelmäksi vestibulaari schwannooman automaattisegmentoinnissa, antaen mahdollisesti kliinistä lisäarvoa. Lisäksi avoimien resurssien hyödyntäminen havaittiin toimivaksi tavaksi rakentaa näitä järjestelmiä.

SYMBOLS AND ABBREVIATIONS

Abbreviations

ceT1	Contrast-enhanced T1 MRI
CNN	Convolutional neural network
DL	Deep learning
hrT2	High resolution T2 MRI
ML	Machine learning
MRI	Magnetic resonance imaging
NIfTI	Neuroimaging Informatics Technology Initiative
ReLU	Rectified linear unit
TCIA	The cancer imaging archive
VS	Vestibular schwannoma

Table of contents

1 Introduction

The field of medical imaging faces challenges in automation and efficiency. While machine learning (ML) offers potential solutions, implementation barriers, such as cost and data constraints persist. This thesis examines the use of deep learning to address these challenges within the specific context of medical computer vision.

ML has rapidly become one of the most important fields in software engineering due to its numerous real-life applications and recent advances in computing power. ML entails learning patterns and predicting outcomes using purpose-built algorithms and statistical models given a set of data. Typically, these algorithms are trained using annotated data and then applied to previously unencountered data for making predictions. Unlike conventional algorithms, machine learning models identify patterns and deduce outcomes without being directly programmed to do so. Neural networks (NNs), a subset of ML algorithms especially capable of detecting patterns, are inspired by the human brain, featuring intertwined layers of 'neurons' that process and relay information. Deep learning (DL), characterised by deep multi-layered NNs, facilitates intricate pattern recognition and is the currently one of the most widely used sub-sets of ML. DL architectures often incorporate convolutional and pooling layers in an attempt to strike a balance between handling large and intricate datasets efficiently while preserving data quality. This type of DL architecture is commonly referred to as a convolutional neural network (CNN). CNNs have been found to be particularly effective in the field of computer vision (Goodfellow et al., 2016; Krizhevsky et al., 2017).

A specific domain that has reaped the benefits of this advancement is medical imaging, where one of the key challenges is the time consuming and difficult nature of image interpretation. CNNs have proven remarkably effective for various medical computer vision tasks (Esteva et al., 2017; Kermany et al., 2018; Ronneberger et al., 2015). Consequently, deep learning systems with the potential to automate and aid in these processes are highly sought after. Modern medical imaging software is no longer limited to visualisation and storage; the incorporation of DL algorithms that aid in image interpretation has been steadily increasing. Despite the growing integration of ML into medical imaging software, many systems only harness a fraction of its potential, leaving significant capabilities underutilised. Leveraging these technologies on a broader scale is still hindered by the high cost associated with specialised hardware required to train these systems and the availability of suitable datasets.

The aim of this thesis is twofold: firstly, to address a challenging problem in medical computer vision by introducing a DL model that automatically segments vestibular schwannoma (VS) from magnetic resonance imaging (MRI); and secondly, to leverage open source computing platforms in conjunction with a publicly available VS dataset (Shapey et al., 2021b, 2021a). Through this approach, this thesis assesses the feasibility and potential of open-source computing resources and data in tackling complex challenges in medical imaging.

VS, also known as acoustic neuroma, is a type of benign intracranial tumour originating from the Schwann cells of the vestibulocochlear nerve in the inner ear. This nerve is responsible for hearing and balance, hence the primary symptoms of VS include ipsilateral hearing loss, balance issues and tinnitus. With steadily rising incidence rates of 3 to 5 cases per 100 000 person-years currently, VS comprises roughly 8% of all intracranial tumours. The rise in incidence is at least partly attributed to a more widespread access to MRIs and improved screening protocols for asymmetric hearing loss (Carlson and Link, 2021). For VS treatment planning and follow-up assessments, the segmentation of the tumour volume from MRI images is important. Typically, this is a cumbersome manual process conducted by a specialist doctor. Given these circumstances, the need for automated systems to aid in VS diagnosis is evident.

In previous studies, DL has shown promise in the automatic segmentation of VS. Multiple studies have found DL algorithms to provide high quality segmentations without human intervention, thus having the potential to improve patient outcomes and facilitate both time and cost savings (Cass et al., 2022; Kujawa et al., 2022; Neve et al., 2022; Shapey et al., 2021c; Wang et al., 2022; Yao et al., 2022). Nevertheless, in this context the accessibility of these technologies has not been thoroughly explored.

The advancements in DL offer transformative solutions to fields such as medical imaging. While the technology holds promise, the widespread adoption of DL may be impeded by the unavailability of data sets and the steep hardware requirements needed for building these systems. This thesis attempts to bridge this gap by leveraging freely available tools to solve a demanding real life medical computer vision problem.

2 Related work

This chapter provides an overview of the previous studies related to the automatic segmentation of VS using DL. These studies demonstrate the feasibility of using such methods for this task. In addition, a public dataset containing MRI scans of VS patients is introduced.

2.1 Open annotated VS dataset

Shapey et al. describe a publicly available dataset consisting of 484 MRIs of 242 adult patients with a single unilateral VS undergoing Gamma Knife Stereotactic Radiosurgery (Shapey et al., 2021b, 2021a). The data includes the contrast-enhanced T1-weighted (ceT1) MRI sequences, the high-resolution T2-weighted (hrT2) MRI sequences, and the manual segmentations used for treatment planning. According to Shapey et al. the provided segmentations were annotated in consensus by the treating neurosurgeon, neuroradiologist and a physicist. Both ceT1 and hrT2 sequences were used for the annotation.

2.2 Deep learning approaches for automated VS segmentation

A key issue in medical computer vision is finding the balance between computational efficiency and sufficient accuracy in the given task, as the datasets used are typically 3D images and therefore substantial in terms of memory requirements. Different approaches to address this problem have been proposed, with most of them relying on pre-processing the image data to a more efficient form. Moreover, the choice between employing a 2D or a 3D CNN architecture becomes a crucial design consideration in alleviating this concern.

Using the dataset described in the previous section, Shapey et al. introduced a 2.5D CNN for automatically segmenting VS from ceT1 and hrT2 (Shapey et al., 2021c). In contrast to a simple 2D or a more computationally demanding 3D approach, the authors implemented a 2.5D U-Net with the intention of gaining insight from the varying in-plane and through-plane resolutions often present in clinical MRI protocols while managing computational overhead. The authors reported Dice scores of 0.93 and 0.88 for ceT1 and hrT2 inputs respectively, showing high segmentation accuracy. Utilising a proportion of the same open dataset, Kujawa et al. proposed a DL model to automatically grade VS from MRI images according to the Koos scale (Kujawa et al., 2022). In its first stage, the grading process involves the automatic segmentation of VS, for which the authors implemented a 3D U-Net. For these intermediate segmentations the authors reported mean Dice scores of 0.94 ± 0.04, 0.91 ± 0.05, 0.94 ± 0.04 for ceT1, hrT2 and ceT1 + hrT2 respectively.

In contrast to the previously described volumetric approaches, Yao et al. describe two 2D U-Nets for segmenting VS from axial slices of ceT1 and hrT2 from 82 VS patients (Yao et al., 2022). Their models achieved Dice scores of 0.89 and 0.85 for the ceT1 and hrT2 based models respectively, demonstrating the feasibility of a more computationally inexpensive approach.

In clinical settings, the bulk of MRI scans typically reveal no abnormalities; however, when training Deep Learning (DL) algorithms for detecting and segmenting tumours, incorporating this reality into the data often risks diminishing the quality of the results. Bridging this gap between current research and practice is essential for implementing these systems to real-world settings. Neve et al. investigated the use of 3D U-Nets in segmenting VS from a dataset of 214 patients with hearing loss that underwent MRI (Neve et al., 2022). VS was found in 134 of these patients leaving a negative sub-group of 80 patients. Based on this data ceT1 and hrT2 based models were built. The sensitivity and specificity for detecting tumours was 100% for the test set. The authors reported a mean Dice score of 0.91 ± 0.10 and 0.87 ± 0.06 for ceT1 and hrT2 respectively. In addition, the ceT1 based model was assessed with the publicly available data set described previously, achieving a mean Dice score of 0.88 ± 0.04.

The significance of extensive and heterogenous datasets in DL is fundamental, particularly in medical imaging, as it enables models to discern intricate patterns, improving the reliability and generalization of the model in complex real-world conditions. Cass et al. explored a combination of a transformer architecture and a 3D CNN for automatic VS segmentation, employing both external and internal datasets containing 105 and 40 MRIs of VS patients, respectively (Cass et al., 2022). The Dice scores reported were 0.87 and 0.82 for the external and internal datasets, respectively. However, when the model, initially trained on the external dataset, was subsequently tested on the internal dataset, there was a substantial decline in performance, with the Dice score decreasing to 0.36, underscoring the need for diverse and substantial training datasets to achieve models that are robust and universally applicable.

Recently, the scope of medical computer vision has broadened to encompass more holistic approaches. In addition to determining the state of the current condition, these techniques attempt to predict, for instance, patient survival and recognise patients that are at a high risk of developing more difficult diseases. Wang et al. examined a DL-based multi-task model for automatic VS segmentation and tumour enlargement prediction on a dataset of ceT1s from 103 VS patients with the goal of identifying the patients at risk for fast tumour growth early on (Wang et al., 2022). The authors implemented a 3D U-Net architecture for the segmentation task with a separate module for the enlargement prediction. They compare the segmentation performance of their multi-task model with a separate 3D U-Net with the same architecture as the segmentation module in the multi-task model and a state-of-the-art nnU-Net model. The median Dice scores for these models were 0.84, 0.83 and 0.86 respectively.

3 Methods

3.1 Data procurement

As described in chapter 2.1. this thesis uses the public VS dataset prepared by Shapey et al. available from The Cancer Imaging Archive (TCIA) (Clark et al., 2013; Shapey et al., 2021b, 2021a). The NBIA data retriever, available from the TCIA, was used to download the data.

3.2 Software requirements

All pre-processing and model training detailed in this chapter were carried out using Python (v. 3.10.11), Nibabel (v. 5.1.0), Scipy (v. 1.10.1), Numpy (v. 1.24.3), sklearn (v. 1.2.2), Segmentation models 3D (v. 1.0.4) and Tensorflow (v. 2.10.1).

3.3 Pre-processing

3.3.1 Organising the data

The downloaded dataset consisted of the ceT1 and hrT2 MRIs for the 242 VS patients in DICOM file format, radiation therapy structures, registration matrices and the contours in JSON format with a combined size of approximately 26 gigabytes. After downloading the image data, it was subsequently organised to a more convenient folder structure matching the directory structure of the contours and registration matrices using the Python script provided by Shapey et al. in the GitHub repository associated with the project (Shapey et al., 2021b). In contrast to the pre-processing guidelines provided by Shapey et al. in their GitHub repository, a Python script was crafted to automatically merge the contours and registration matrices to this new folder structure instead of manually merging them.

3.3.2 Preparing the images

The DICOM file format is a common image format used in the medical domain. It stores 3D images in a series of files, each file representing a single 2D slice of the 3D image. For DL purposes this format can be inconvenient, therefore a Python script provided in the GitHub repository of Shapey et al. was utilised to transform all DICOM files and JSON contours into Neuroimaging Informatics Technology Initiative (NIfTI) file format, which has the advantages of containing the volumetric data in a single file along with the compatibility for common Python libraries facilitating a more efficient pre-processing pipeline. For the purpose of this thesis, ceT1 MRI sequences were chosen for model training, as they were predominantly used for the manual tumour segmentation (Shapey et al., 2021b). A Python script was crafted for extracting the ceT1 sequences and corresponding tumour masks.

The in-plane image dimensions for the original data were 512 x 512 pixels with a resolution of 0.4 × 0.4 mm. The depth was 120 image slices for 224 patients and 160 slices for 18 patients with a slice thickness of 1.0-1.5mm. To accommodate larger batch sizes and facilitate a more efficient training process the images were downsampled to the in-plane dimensions of 64 x 64 pixels with bicubic interpolation. Similarly, the masks were downsampled to the same in-plane dimensions using nearest neighbour interpolation to avoid any unwanted artifacts. The depth dimension for the images and corresponding masks was subsequently determined by choosing 64 consecutive slices around the centre most mask slice containing annotated structures. The images and masks were then saved as Numpy arrays resulting in two arrays with dimensions of 242 x 64 x 64 x 64, where the first axis represents the number of patients and the axes 1-3 represent height, width and depth respectively. Due to a varying range of voxel intensities, the images were then normalised per 3D image using Z-score normalisation.

3.4 Deep learning model

A 3D U-Net inspired model was chosen for the segmentation task due to the architecture's excellent performance in biomedical image segmentation in the past (Çiçek et al., 2016; Ronneberger et al., 2015). A simple attention mechanism was implemented to enhance the model's capability of focusing on the relatively small area of interest, as this mechanism has been shown to increase performance in previous work (Oktay et al., 2018). Residual blocks were used to enable a more efficient gradient flow. The model was built using the Keras API of the Tensorflow framework. A schematic representation of the model is depicted in Fig.1.

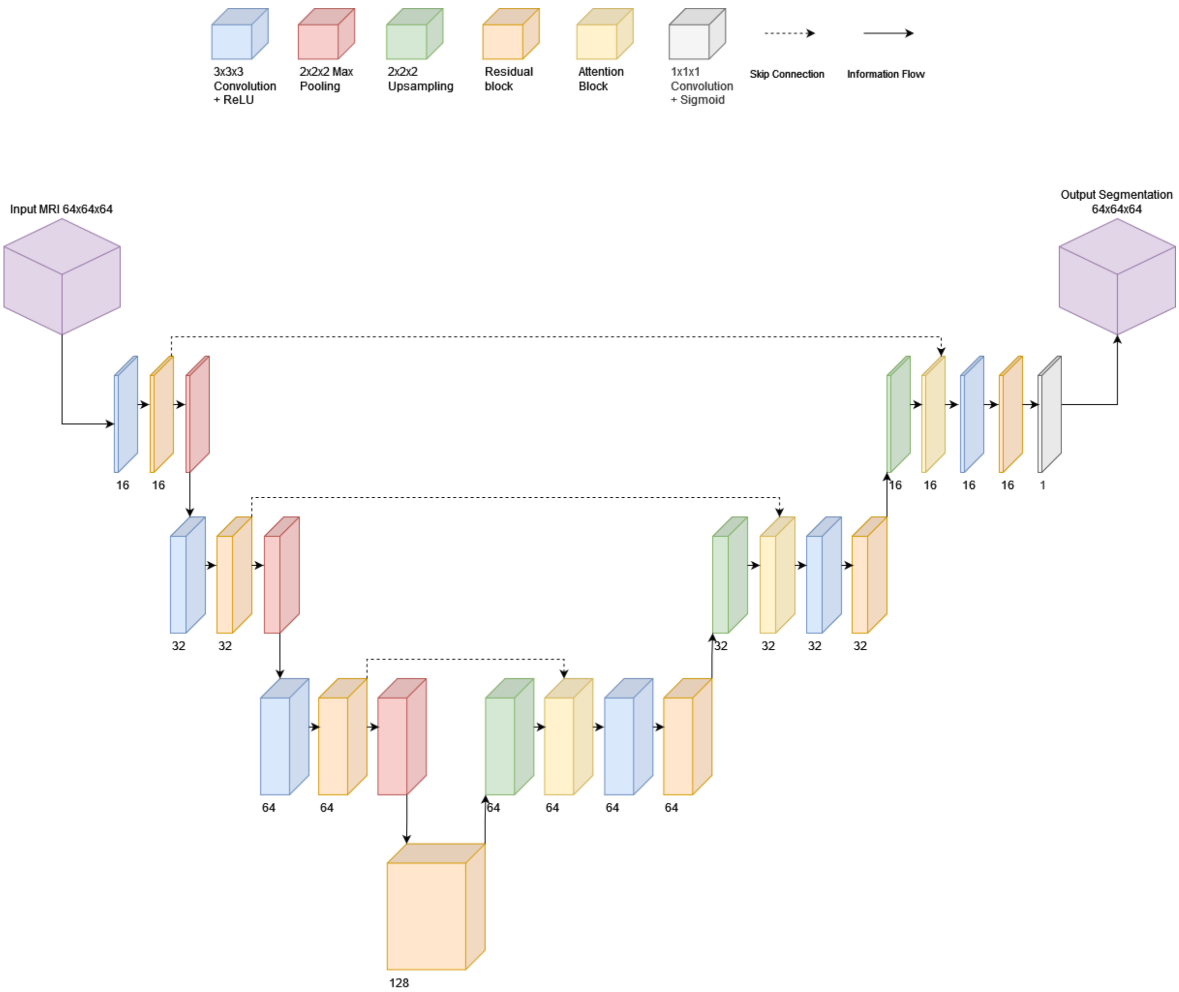

Figure 1. Model architecture.

In the model schema, the number of filters is represented as an integer below each block. The residual blocks consist of two 3D convolutional operations with a kernel size of 3x3x3, a batch normalisation layer, Rectified Linear Unit (ReLU) activation and a shortcut path with a 3D convolutional layer with a kernel of size 1x1x1. These paths are then concatenated and passed forward. The attention blocks take in the upsampled output of the previous layer together with the skip connection input from the encoder path at the corresponding level. These are then added and passed through ReLU activation. The attention coefficients are then extracted using 1x1x1 convolution and sigmoid activation. The final output of the block is the original feature map from the skip connection multiplied by the coefficients. The output of the attention block is then further concatenated with the skip connection from the encoder path and passed through a 3x3x3 convolution and a residual block before being forwarded to next layer. To obtain the final segmentation map, the last layer of the model consists of a simple 1x1x1 convolution with a single filter to reduce the feature map dimensions from 64x6x64x16 to 64x64x64x1 which is the final output.

3.5 Model training and evaluation

The dataset was first divided into training and test subsets with an 80/20 ratio. To ensure a robust training procedure with a relatively limited dataset, 5-fold cross-validation was used. Within the cross-validation loop, a single model was trained for a maximum of 100 epochs with early stopping enabled using a patience value of 15 epochs. A dual loss function combining Dice loss and binary cross-entropy from the segmentation models 3D library was used for training the model (Iakubovskii, 2019). The Adam optimiser was used to minimise the loss during training with an initial learning rate of 0.0001 and an exponential decay rate of 0.96 with 100000 steps. Dice score was used as the metric during training and evaluation and is calculated as follows:

Let A be the predicted binary mask and B the ground truth binary mask.

$$D(A,B) = \frac{2 * |A \cap B|}{|A| + |B|}$$

Where $|A \cap B|$ depicts the intersection of the A and B, and $|A|$ and $|B|$ are the number of pixels with value 1 in A and B respectively.

After each fold weights from the epoch with the highest Dice score for the validation set were restored and the model was saved. In addition, at the end of each fold the trained model was asked to give segmentations of each patient of the corresponding validation set, which were then evaluated with the Dice score. After the cross-validation the model with the lowest standard deviation in terms of Dice score for the patients in the validation set was subsequently chosen to be trained further utilising the whole training subset. The maximum number of training epochs for this model was set to 50. To avoid overfitting, early stopping was utilised with a patience value of 5 epochs. The weights of the epoch with the highest Dice score were restored after training.

4 Results

During cross-validation all models converged towards the local optima reasonably fast in roughly 20 epochs with a modest amount of overfitting. The Dice score as a function of elapsed epochs for training and validation sets is depicted in Fig.2.

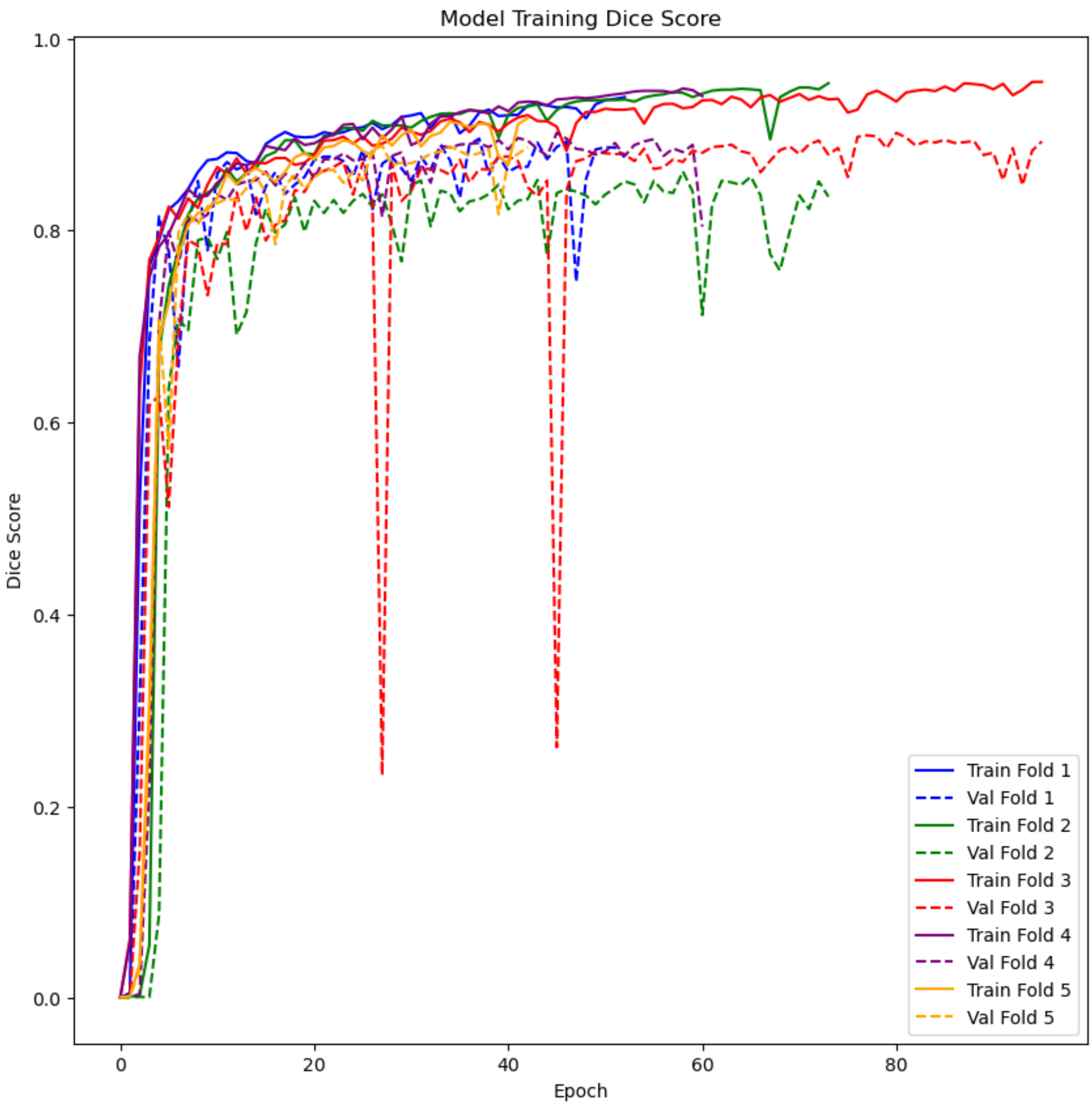

Figure 2. Dice scores for each training and validation fold as a function of elapsed epochs.

The average Dice scores for a given validation fold ranged from 0.84 to 0.90 with standard deviations in the range of 0.07 to 0.11 excluding the second fold, where the standard deviation was 0.21. The difference in the validation performance of the second fold is observable in Fig.1, where the overfitting is greater in comparison to the other folds. The model chosen for further training based on the standard deviation was then trained for a total of 14 epochs achieving a Dice score of 0.94 on the training set. This model was then evaluated on the independent test set. The model reached a mean Dice score of 0.85 with a standard deviation of 0.21. These results are detailed in Table 1.

Table 1. Mean Dice scores and standard deviations of the cross-validation folds and the independent test set.

Test data	Dice Score ± Standard Deviation
Validation fold 1	0.89 ± 0.11
Validation fold 2	0.84 ± 0.21
Validation fold 3	0.89 ± 0.09
Validation fold 4	0.90 ± 0.07
Validation fold 5	0.88 ± 0.09
Test set	0.85 ± 0.21

On average, the model performed well on most test patients. However, for two test cases it failed to segment any relevant structures. These cases were indices 33 and 231 of the test set corresponding to indices 131 and 8 in the original dataset respectively. The model performance in terms of the Dice score for the test patients is depicted as a box plot in Fig. 3, where Interquartile Range (IQR) is the difference between the third quartile (Q3) and the first quartile (Q1). Inner fences were calculated as:

Lower inner fence: Q1-1.5xIQRQ1-1.5xIQR

Upper inner fence: Q3+1.5xIQRQ3+1.5xIQR

Data points outside of the fences were considered outliers.

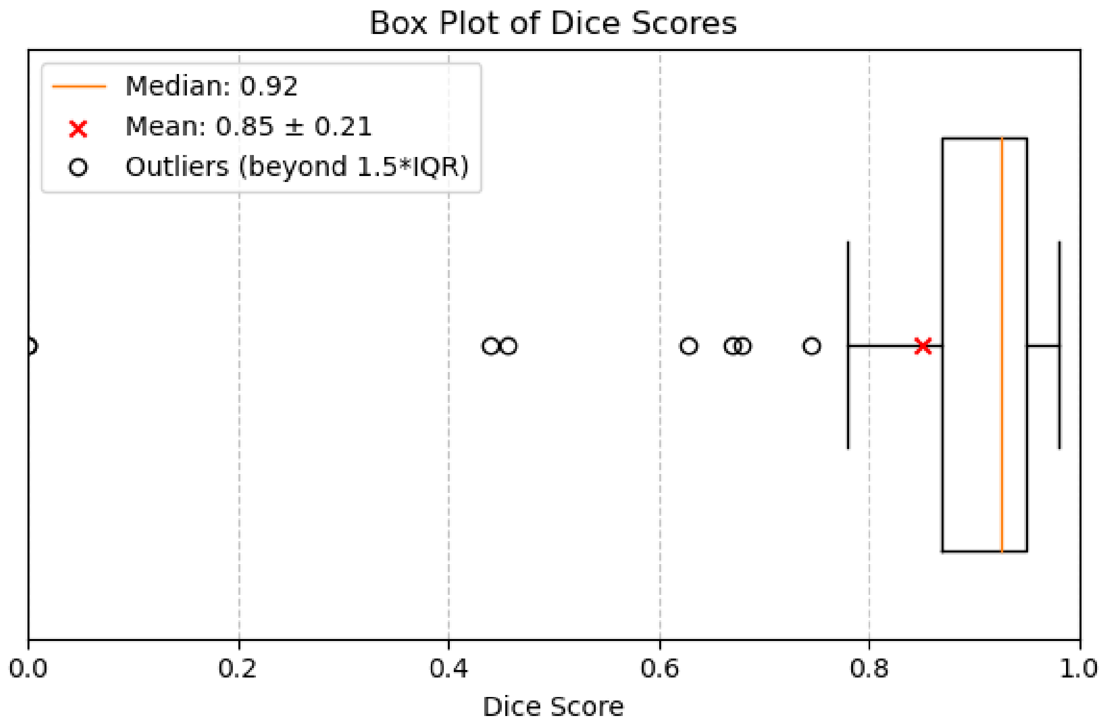

Figure 3. Box plot of the predictions on the test set.

An example of a successful segmentation is demonstrated in Fig. 4, which depicts the test case 19 (number 118 in the original dataset) with a Dice score of 0.97. The blue contours represent the ground truth masks and the red contours represent the predicted mask.

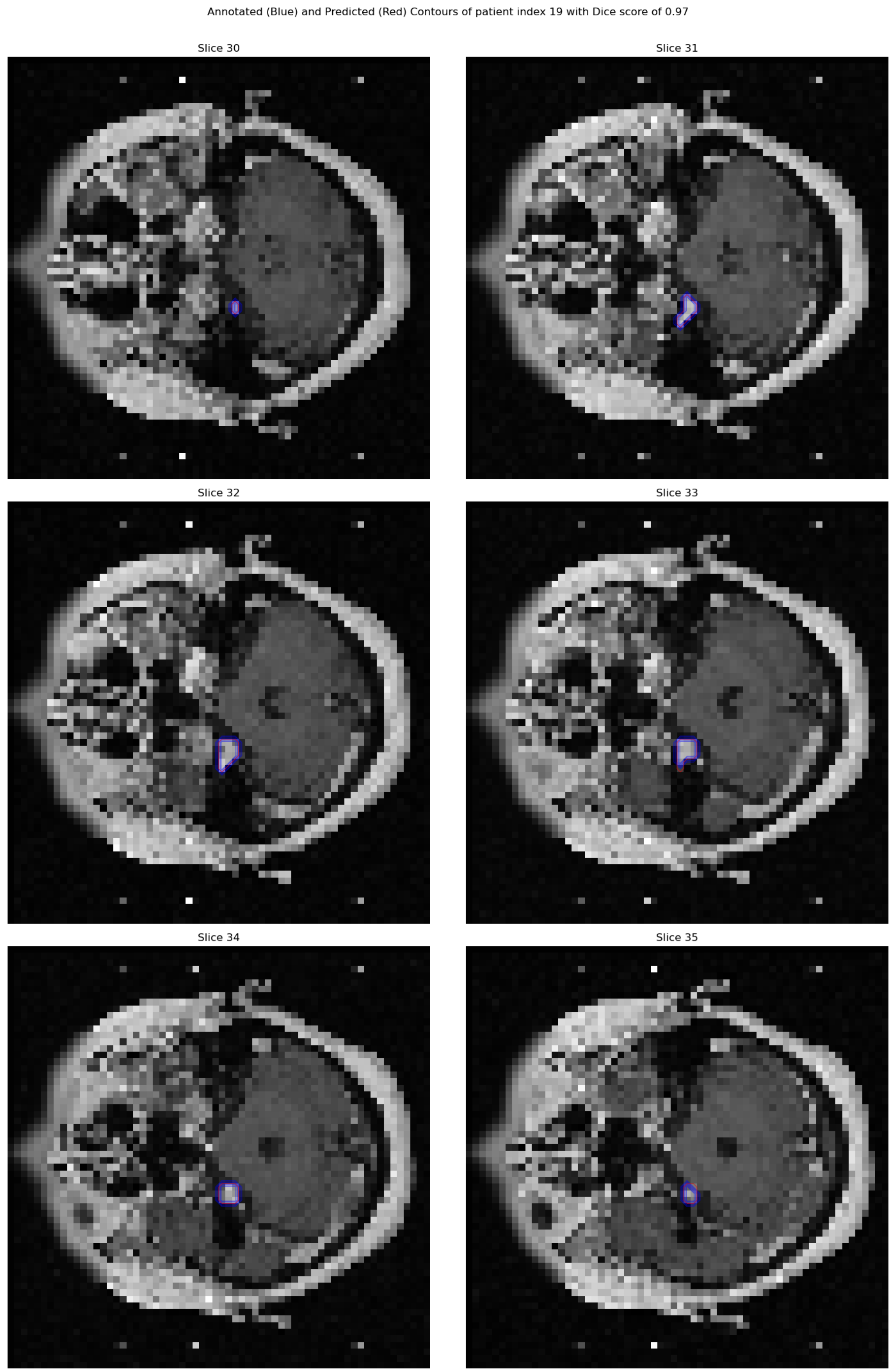

Figure 4. Example of a successful segmentation result.

5 Discussion

The developed DL model shows promise in the automatic segmentation of VS from MRIs producing high quality segmentations with a Dice score of 0.85 ± 0.21 for the test set – a result resembling the Dice score of 88 ± 0.04 reported by Neve et al. using a proportion of the same ceT1 data, excluding 47 post surgery patients, as an external test set (Neve et al., 2022). Similarly, Shapey et al. describe a model with a slightly higher mean Dice score of 0.93 ± 0.04 for the same subset of ceT1 images of the dataset (Shapey et al., 2021c). In addition, Kujawa et al. introduced a method which achieved a mean Dice score of 0.94 ± 0.04 for ceT1 images using a dataset of 308 VS patients containing 188 of the same patients as the previous studies (Kujawa et al., 2022). However, some limiting factors for the direct comparison of these methods need to be considered. Firstly, the methods described by Neve et al. and Kujawa et al. only share a proportion of the data with the method proposed in this thesis. Secondly, Shapey et. al describe a 2.5D architecture, whereas the others utilise a 3D-based approach. Moreover, Kujawa et al. utilised data augmentation to increase the size of their dataset, a common method for dealing with limited data sets in DL, which was not utilised in this thesis nor was it described by Neve et al. or Shapey et al. The authors also report the use of ensemble methods, where the segmentation is produced by a majority vote based on predictions of multiple models, rather than just one model as described in this thesis. Furthermore, for the proposed method in this thesis, significant image resampling was carried out to address the lack of computational resources, which can hinder performance. These factors together with a different model architecture in this thesis might also, at least partly, account for the larger standard deviation seen with the proposed method.

A crucial aspect of this work was to explore the feasibility of using freely available computing resources and data to address DL problems with possible real-world impact. For the purposes of medical computer vision, TCIA provides a large archive of anonymised high quality medical imaging datasets. The data used in this study, provided by Shapey et al., is a fine example of this and was found to facilitate a quick and relatively easy access to a DL-ready dataset (Shapey et al., 2021b). The experiments described in this thesis are publicly available in the GitHub repository and can be replicated in Google Colab following the instructions provided in the README-section of the repository, thus demonstrating the feasibility of using openly sourced resources for real-world DL projects. However, it should be noted that the free version of Colab can be sensitive in reallocating resources it identifies as idle – a feature that was found bothersome during the more lengthier trial runs of the 5-fold cross-validation and were therefore carried out on a local machine.

The possible clinical implications of an accurate automatic segmentation procedure for tumours in the field of radiology are significant. DL assisted segmentation workflows will greatly benefit from improved efficiency in terms of time and cost, which in turn can lead to better patient outcomes for instance by reducing time needed for diagnosis or radiation therapy planning. The proposed method has an inference time of circa 20 milliseconds for a single MRI volume when run on a GeForce RTX 3070 GPU, underscoring the difference in speed between a DL model and a human observer. In addition, manual segmentations often suffer from inherent intra- and interobserver variability, which can be mitigated using automated systems.

However, some limitations of this study need to be addressed. The number of patients included in the dataset, although substantial for a medical dataset, is still limited in terms of typical datasets used in DL. The lack of heterogeneity in the data might cause the model to overfit to a sample that does not accurately capture the true distribution in a real-world setting, thereby diminishing the model's capability of generalising to a wider population. Moreover, only Dice score was used as the evaluation metric. While it is an efficient method for measuring spatial overlap and a straightforward benchmark for previous work, more descriptive metrics might be required for gaining clinical insights.

In future work several strategies can be implemented to ensure a model robustly generalises across diverse patient groups and MRI sequences. Primarily, acquiring an extensive dataset coupled with data augmentation techniques. Subsequently, leveraging more sophisticated and/or pre-trained models through transfer learning can be beneficial. Lastly, employing ensemble models serves as another viable approach. For a more comprehensive evaluation of the segmentation performance, surface distance metrics such as Hausdorff distance could be implemented.

6 Conclusions

This thesis demonstrates that solving complex DL problems using open-source data and resources is feasible and yields promising results. The results achieved with the proposed method in VS segmentation resemble those described in previous literature, thus reinforcing the consensus that DL has great potential to aid in the diagnosis and treatment of these patients (Kujawa et al., 2022; Neve et al., 2022; Shapey et al., 2021c).

When incorporated into clinical medical image analysis, DL systems will reduce the time it takes to interpret images accurately, thus providing significant time and cost savings and alleviating the clinician's workload. Furthermore, by reducing the variability between interpretations, the integration of these systems to existing workflows will facilitate the standardisation of treatment and follow-up protocols by eliminating some of the subjectivity inherent in human observers. As a result, patients will receive treatment in a more timely and precise manner, ultimately leading to better outcomes.

To further facilitate the broad scale research and implementation of DL systems using open-source methods, more emphasis needs to be placed on the availability of quality datasets together with the usability of freely available computing platforms for the hardware intensive training procedures. The increased availability of suitable datasets will benefit the field of medical computer vision, as the generalisability of the models is closely correlated with the amount of available training samples, thereby incentivising clinical integration. Integration into existing image analysis software could also help address the data availability problem, if implemented in a way that allows for continuous integration and delivery.

Future research and development of these systems should be conducted in close collaboration with radiologists and other medical professionals, as subject matter expertise is key with VS segmentation and other similar highly domain specific DL problems. Given the non-standard and potentially atypical nature of these data, a data centric, in contrast with model centric, development approach might be beneficial. Going forward, collaboration with subject matter experts can also aid in increasing model interpretability, an important consideration in safety-critical software development domains.

References

Carlson, M.L., Link, M.J., 2021. Vestibular Schwannomas. N. Engl. J. Med. 384, 1335–1348. https://doi.org/10.1056/NEJMra2020394

Cass, N.D., Lindquist, N.R., Zhu, Q., Li, H., Oguz, I., Tawfik, K.O., 2022. Machine Learning for Automated Calculation of Vestibular Schwannoma Volumes. Otol. Neurotol. Off. Publ. Am. Otol. Soc. Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol. 43, 1252–1256. https://doi.org/10.1097/MAO.0000000000003687

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Springer International Publishing, Cham, pp. 424–432.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J. Digit. Imaging 26, 1045–1057. https://doi.org/10.1007/s10278-013-9622-7

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. https://doi.org/10.1038/nature21056

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Iakubovskii, P., 2019. Segmentation Models [WWW Document]. URL https://github.com/qubvel/segmentation_models

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, Justin, Prasadha, M.K., Pei, J., Ting, M.Y.L., Zhu, J., Li, C., Hewett, S., Dong, Jason, Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A.N., Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K., 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 172, 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet Classification with Deep Convolutional Neural Networks. Commun ACM 60, 84–90. https://doi.org/10.1145/3065386

Kujawa, A., Dorent, R., Connor, S., Oviedova, A., Okasha, M., Grishchuk, D., Ourselin, S., Paddick, I., Kitchen, N., Vercauteren, T., Shapey, J., 2022. Automated Koos Classification of Vestibular Schwannoma. Front. Radiol. 2, 837191. https://doi.org/10.3389/fradi.2022.837191

Neve, O.M., Chen, Y., Tao, Q., Romeijn, S.R., de Boer, N.P., Grootjans, W., Kruit, M.C., Lelieveldt, B.P.F., Jansen, J.C., Hensen, E.F., Verbist, B.M., Staring, M., 2022. Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium-based Contrast Material: A Multicenter, Multivendor Study. Radiol. Artif. Intell. 4, e210300. https://doi.org/10.1148/ryai.210300

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., others, 2018. Attention u-net: Learning where to look for the pancreas. ArXiv Prepr. ArXiv180403999.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Shapey, J., Kujawa, A., Dorent, R., Wang, G., Bisdas, S., Dimitriadis, A., Grishchuck, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S., Ourselin, S., Vercauteren, T., 2021a. Segmentation of Vestibular Schwannoma from Magnetic Resonance Imaging: An Open Annotated Dataset and Baseline Algorithm (Vestibular-Schwannoma-SEG). https://doi.org/10.7937/TCIA.9YTJ-5Q73

Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., Grishchuk, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S.R., Bisdas, S., Ourselin, S., Vercauteren, T., 2021b. Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm. Sci. Data 8, 286. https://doi.org/10.1038/s41597-021-01064-w

Shapey, J., Wang, G., Dorent, R., Dimitriadis, A., Li, W., Paddick, I., Kitchen, N., Bisdas, S., Saeed, S.R., Ourselin, S., Bradford, R., Vercauteren, T., 2021c. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI. J. Neurosurg. JNS 134, 171–179. https://doi.org/10.3171/2019.9.JNS191949

Wang, K., George-Jones, N.A., Chen, L., Hunter, J.B., Wang, J., 2022. Joint Vestibular Schwannoma Enlargement Prediction and Segmentation Using a Deep Multi-task Model. The Laryngoscope. https://doi.org/10.1002/lary.30516

Yao, P., Shavit, S.S., Shin, J., Selesnick, S., Phillips, C.D., Strauss, S.B., 2022. Segmentation of Vestibular Schwannomas on Postoperative Gadolinium-Enhanced T1-Weighted and Noncontrast T2-Weighted Magnetic Resonance Imaging Using Deep Learning. Otol. Neurotol. Off. Publ. Am. Otol. Soc. Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol. 43, 1227–1239. https://doi.org/10.1097/MAO.0000000000003693