Volume-based Performance not Guaranteed by Promising Patch-based Results in Medical Imaging

Abhishek Moturu
University of Toronto[1]
moturuab@cs.toronto.edu

Sayali Joshi
The Hospital for Sick Children[3]
sayali.joshi@sickkids.ca

Andrea Doria
The Hospital for Sick Children[3]
andrea.doria@sickkids.ca

Anna Goldenberg
University of Toronto[12]
anna.goldenberg@utoronto.ca

[1]Department of Computer Science [2]Department of Laboratory Medicine and Pathobiology [3]Department of Diagnostic Imaging

Abstract

Whole-body MRIs are commonly used to screen for early signs of cancer. In addition to the small size of tumours at onset, variations in individuals, tumour types, and MRI machines increase the difficulty of finding tumours in these scans. Using patches, rather than whole-body scans, to train a deep-learning-based segmentation model with a custom compound patch loss function, several augmentations, and additional synthetically generated training data to identify areas where there is a high probability of a tumour provided promising results at the patch-level. However, applying the patch-based model to the entire volume did not yield great results despite all of the state-of-the-art improvements, with over 50% of the tumour sections in the dataset being missed. Our work highlights the discrepancy between the commonly used patch-based analysis and the overall performance on the whole image and the importance of focusing on the metrics relevant to the ultimate user − in our case, the clinician. Much work remains to be done to bring state-of-the-art segmentation to the clinical practice of cancer screening.

1 Introduction

Finding small anomalies in large images is challenging for many reasons. Especially within medical imaging, the small relative sizes of the regions of interest, lack of large amounts of data, paucity of positive cases (anomalies such as cancer), heterogeneity in the data and the anomalies, and variation in expert labels exacerbate the effectiveness of machine learning approaches in detecting anomalies. Patch-based approaches are commonly used in the field to boost the success of detection [1, 2, 3, 4].

In pediatric whole-body MRIs (wbMRIs), early detection of cancers at onset (i.e. at a very small size within the body) is crucial for a good prognosis, especially in children with cancer predisposition syndromes [5]. Variations in the child, the type of cancer, and the type of MRI must be considered during the evaluation of wbMRIs. A screening tool that can be used by radiologists and oncologists with varying expertise to assist them in finding regions where there is high likelihood of a tumour can save time and improve tumour detection capabilities.

Deep-learning-based segmentation approaches have seen great advancements in recent years, especially within the field of medical imaging. In this paper, we present our modified patch-based approach for tumour segmentation to overcome the challenges with having few positive cases (i.e. cancers) and vast data heterogeneity. Even after attempting to address these issues (see Table 1 in Section 2) by splitting the data into patches, using the U-Net model, training with asymmetric unified focal loss, dice loss, contour perimeter loss, and contour difference loss (see Section 4), selecting different proportions of positive and negative cases, and utilizing data augmentations and synthetic data, there is still room for improvement in the tumour segmentation performance in wbMRIs.

I Can't Believe It's Not Better Workshop at NeurIPS 2022.

Our results indicate the need to consider several metrics to understand how patch-based performance and overall performance affect each other and that given a set of patches related to a larger task, the patch-based performance may not be fully indicative of how the overall performance may turn out.

2 Related Work

There has been a lot of work in recent years in tumour detection and localization in MRIs, however the amount of work in pediatric whole-body MRIs focused on finding tumours at onset in any part of the body continues to be rare. Non-segmentation-based deep learning techniques for tumour detection and localization include classification [6, 7, 8, 9], anomaly detection with generative adversarial networks [10, 11, 12, 13, 14] or variational auto-encoders [15, 16, 17], or a combination of these approaches [18, 19]. Segmentation-based deep learning techniques for tumour detection and localization [20, 21, 22] perform well within organ-specific domains due to the well-defined nature of the body regions and anomalies (i.e tumours). However, all of these methods have trouble in domains with very few positive cases and vast data heterogeneity, especially when the size of the anomalies is very small within the data. Patch-based techniques can incorporate any of the above methods [1, 2, 23, 24, 3, 4, 25, 26, 27], but aggregating patch-based results and metrics for meaningful and effective overall results is not always trivial [28].

Problem	Attempted Solution	Outcome
volumes are 3D	3D models (varying slice thickness), work with slices and 2D models [29]	+ larger relative tumour size with 2D slices
volumes are very large	work with patches [2]	+ larger relative tumour size with patches
tumours are very rare	generative modeling [11, 15]	− difficult due to low-data and data heterogeneity
tumours are very small	work with patches [1, 3], segmentation [20, 21, 22]	+ works better in low-data domains
varying bands of brightness in wbMRIs	contrast-limited adaptive histogram equalization [30]	+ better contrast and tumour visibility
some volumes are noisy	denoising techniques [31]	+ improved clarity and contrast
varying positions	registration [32]	− difficult due to data heterogeneity
varying age and sex	conditioning on age and sex [33]	− slower and more unstable training
varying size, shape, and location of tumours	augmentations [34], add synthetic tumours to training set [23, 24]	+ increased diversity of training data

Table 1: The problems encountered, the attempted solutions, and their + or − outcomes.

3 Dataset

Our dataset contains 675 wbMRI volumes (coronal STIR) from the imaging database of The Hospital for Sick Children in Toronto. On average, there are around 40 slices per volume, since the thickness of the slices varies across MRIs based on the MRI machine and settings. The number of slices ranges from 25 to 50 slices per volume. The dimensions of each slice also vary based on the MRI settings and the age, size, and height of the child, but on average, they are $\sim$3000 pixels $\times$ $\sim$800 pixels.

Only 27 volumes contain tumours and there are a total of 60 tumours. Out of the 27,437 slices in the dataset, only 120 slices contain tumour sections, with the number of slices spanning a tumour ranging from 1 to 7 slices. The approximate diameter of each tumour ranges from 10 pixels to 100 pixels, which is very small compared to the size of a whole-body MRI volume.

After de-identification was used to remove all confidential information from the wbMRIs, the corresponding radiology reports, which indicated cancer locations within the body, were used by our radiologist fellows to perform manual annotations (i.e. segmentations) of tumours using a segmentation tool. These were the masks that were used to train our tumour segmentation model.

We split the volumes into 4 folds for 4-fold cross-validation, with approximately the same number of slices with tumours in each, and also leave out a test set of volumes to evaluate performance.

3.1 Patches and Augmentations

As discussed before, due to the vast variation in tumours and wbMRIs and the paucity of tumours across the dataset, we relied on 2D patches and augmentations to train our model to find tumours efficiently and effectively. First, we find the number of tumours in each segmentation mask label along with their locations by finding the total number of contours. Then, for each tumour, we take a patch with the tumour at its centre and employ combinations of the following augmentations:

- Take multiple patches containing tumours with the centre uniformly randomly moved around by a few pixels in the x- and y- dimensions in a given slice to be able to find tumours that are not at the exact centre of a given test patch.
- Take multiple patches containing tumours with small zooms in or out of the slice to be able to find tumours occupying various proportions of a given test patch.

Then, we take patches without tumours in the following manner:

- Take multiple patches from anywhere within the volume to introduce patches with no tumours (i.e. negative samples) into the dataset.
- Take multiple patches from brighter parts of the volume (brain, bones, and parts of the abdomen) to avoid the model from learning all bright spots as tumours. This is done automatically by randomly taking patches with above a certain average brightness threshold.
- Take multiple patches from the same approximate location in the current volume in several different volumes that do not have a tumour in that location so that the model can learn to identify parts of the body with and without tumours more effectively.

Then, for 25% of all of the patches, we employ the following augmentations:

- Take multiple patches with the entire patch multiplied by a value smaller than and close to 1 for invariance in small changes to the brightness.
- Take multiple patches with the each pixel in a patch multiplied by a value smaller than and close to 1 for invariance to small amounts of noise.

We applied these augmentations in conjunction with each other, rather than separately, to increase the variation in the location, shape, texture, and size (relative to the body and relative to the patch) of the tumour and allow the model to focus on learning to find tumours while remaining invariant to these factors. Note that the patch from the volume and its corresponding mask are taken from exactly the same location. All patches are 128 pixels $\times$ 128 pixels or resized to those dimensions in the case of the small zoom in and zoom out augmentations.

The patches are obtained from the cross-validation set and split by volume. We use 5707 patches with a 70%-30% training/validation split. 25% of the training and validation sets contain tumours.

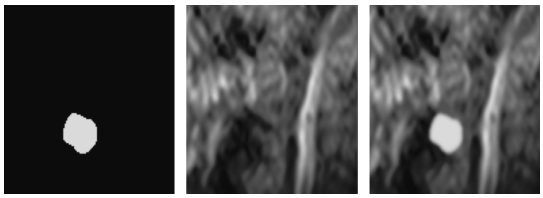

Figure 1: The generation and insertion of a synthetic tumour as described below.

3.2 Synthetic Tumour Generation

To increase the prevalence of positive cases (tumours) within the dataset of patches, we use the following process to generate patches with synthetic tumours (see Figure 1):

1. Create a tumour mask of random size by generating a random number of points, using Graham's scan [35] to find the convex hull, and interpolating smoothly between the points.
2. Choose a patch from within the volume without any tumours.
3. Choose a random brightness for the generated tumour and blend the generated tumour mask from step 1 onto the patch chosen in step 2 to match the texture.

This method allows us to generate synthetic tumours of varying shape, size, brightness, and location within the body to augment the training dataset.

4 Model and Loss Function

We use the U-Net model [20] along with a loss function that includes asymmetric unified focal loss [36], Dice loss [37], contour perimeter loss [38], and contour difference loss. The coefficients of each of these loss functions in the total patch loss (see Equation 1) are tunable hyperparameters.

$$\begin{aligned} Loss_{Patch} = \alpha * Loss_{AsymmetricUnifiedFocal} \\ + \beta * Loss_{Dice} \\ + \gamma * Loss_{ContourPerimeter} \\ + \delta * Loss_{ContourDifference} \end{aligned} \tag{1}$$

Each of the four parts of the patch loss function (Equation 1) serve a specific purpose:

- $Loss_{AsymmetricUnifiedFocal}$ as defined in [36] is based on the combination of:
  - asymmetric Focal Tversky loss, which is similar to Dice loss but assigns weights to false positives and false negatives to better balance of precision and recall (enhances the hard class, suppresses the easy class)
  - asymmetric Focal loss, which is similar to cross-entropy loss but works better with imbalanced sets (suppresses the hard class, enhances the easy class)
- $Loss_{Dice}$ as defined in [37] is the pixel-based F1-score, which is the harmonic mean between precision and recall (to deal with the imbalance between foreground and background pixels)
- $Loss_{ContourPerimeter} = \left(\sum_{p\in\Omega} \widehat{y_p^F} - \sum_{p\in\Omega} y_p^F\right)^2$ as defined in [38] measures the difference between the perimeter of the true tumour and the predicted tumour to ensure that the model does not over-guess or under-guess the size of a tumour.
- $Loss_{ContourDifference} = \left(\sum_{p\in\Omega} (\widehat{y_p^F} - y_p^F)\right)^2$, which we introduce based on the previous loss, measures the number of the pixels where the perimeter of the true tumour and the predicted tumour do not overlap to ensure that the model properly localizes tumour borders.

For the contour-based losses, note that $\Omega \subset \mathbb{R}^2$ is the spatial image domain and y_p^F (or $\widehat{y_p^F}$) is the true (or predicted) value of pixel p if it belongs to the true (or predicted) contour and 0 otherwise $-$ F is the function that extracts the contour from an image, as detailed in [38].

The number of patches and augmentations (discussed in Section 3.1), the batch size (16), the optimizer (Adam [39]), the learning rate ($1e-5$), and the learning rate scheduler (ReduceLROnPlateau) are the other tunable hyperparameters.

5 Experiments

Model	Dice Score	TPR	FPR	TPR > 10%
U-Net w/ $Loss_{Dice}$	0.9097 / **0.8200**	0.9874 / 0.9391	0.0080 / **0.0096**	0.9926 / 0.9550
U-Net w/ $Loss_{Patch}$	0.8328 / 0.7644	0.9834 / 0.9457	0.0179 / 0.0139	0.9908 / 0.9651
U-Net w/ $Loss_{Patch}$ + Augmentations + Synthetic Tumours	0.8014 / 0.7577	0.9754 / **0.9527**	0.0167 / 0.0155	0.9858 / **0.9704**

Table 2: Average of the patch-based results across the folds for training / validation. TPR > 10% signifies the percent of patches that matched the predicted mask with the true mask by over 10% of the size of the true mask. True negative patches count towards this number.

We take the best trained patch-based model from Table 2, i.e. the U-Net model trained with $Loss_{Patch}$ + Augmentations + Synthetic Tumours, and use a sliding window approach with 25% overlap to get a prediction mask for each slice in every volume of every fold using MONAI [40]. The whole-body results on the cross-validation set for the volumes with tumours are as follows: 27/53 tumour detections in 16/21 volumes. For the test set, we have: 6/7 tumour detections in 5/6 volumes.

The false-positive rate in a given volume ranges from 0.5% to 2% of the pixels, but tumours, if they exist, account for approximately 0.01% to 0.0001% of a given volume, which means that there are 50-20,000 times more false-positive pixels than true-positive pixels. False positive pixels usually occur in the brighter parts of the body.

Although it is important for both the false-positive rate and the false-negative rate to be low and there is a trade-off in trying to minimize both, it is more important to not miss too many tumours (i.e. lowering the false-negative rate is a bit more crucial). However, if it is at the expense of a very high false positive rate, then there will be too many areas highlighted with high tumour probability, which would make it challenging to differentiate between the true positives and the false positives to find the actual tumours. It is also important to note that the Dice score is easily affected by the size of the tumours and it is therefore more insightful to consider a variety of metrics to evaluate patch-level performance. See Figure 2 to see examples of patch-level performance and Figure 3 to see examples of performance on entire slices.

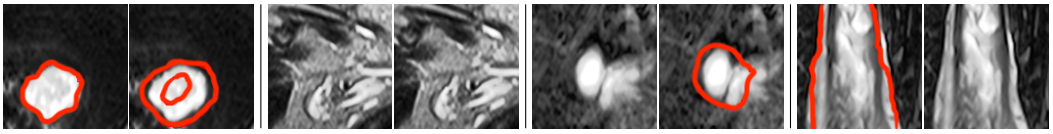

Figure 2: Examples of a true positive, true negative, false positive, and false negative when identifying a tumour (from left to right). Each pair of images contain the contours of the true and predicted masks, respectively (red outline).

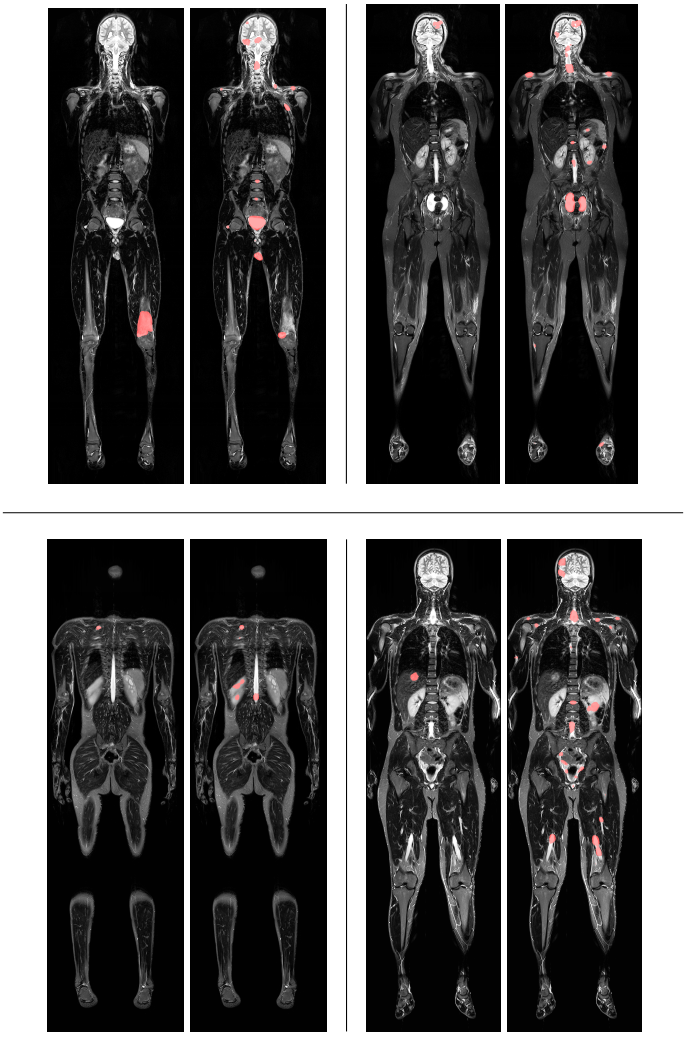

Figure 3: Four examples of whole-body MRI slices in pairs of true mask overlays and predicted mask overlays, respectively. Top left: part of the knee tumour is identified. Top right: entire brain tumour is identified. Bottom left: tumour above lung is identified. Bottom right: tumour in the liver is missed. False positives exist in every example, usually in the brighter parts of the image.

For each of the 120 tumour sections in the cross-validation set, we compare the average brightness and overall size of the tumour in Figure 4 when a tumour is detected or missed. Note that a tumour is classified as a "detect" when there is any overlap between the ground truth label and the prediction of the segmentation model and as a "miss" otherwise. There are no clear patterns, but we can notice that very large and very bright tumours are usually detected, whereas very small and very faint tumours are usually missed by our segmentation model. A total of 58/120 tumour sections are detected, which means that just more than half of the tumour sections are missed.

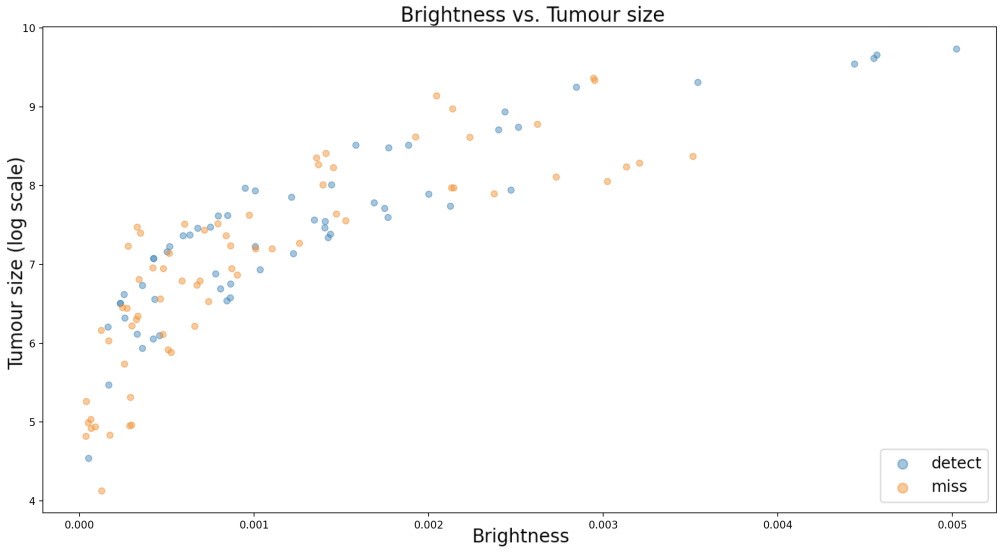

Figure 4: Average tumour brightness vs. tumour size (log scale) when a tumour is detected or missed.

The discrepancy between the results in the 3rd row of Table 2 and Figure 4 suggest that strong performance at the patch-level does not necessarily translate to the whole-body-level due to the fact that we curate the dataset of patches, have a much higher imbalance at the whole-body-level, and may be unable to properly foresee the false-positive rate at the whole-body-level when only considering the patch-level performance.

For example, the following approaches improved patch-based results but worsened the whole-body-based results by either increasing the false-positive rate or the false-negative rate: more proportion of augmentations, more proportion of patches with synthetic tumours, and more weighting in the loss function to the parts which deal with lowering the false-positive rate (i.e. asymmetric unified focal loss and Dice loss).

6 Discussion & Future Work

We employ many tools to make use of the small amount of positive cases and large amount of negative cases. We split the data into patches, use the U-Net model, select the appropriate proportions and types of patches, utilize a compound loss function consisting of losses dealing with each issue in the domain, augment large portions of the data, and add patches with synthetic tumours. Although each of these approaches consistently improved the results at the patch-level from the baseline, we noticed that the whole-body-level results did not see major improvements and, in some cases, worsened.

Our work highlights the need for future methods to follow through on finding a more holistic approach to consolidate strong patch-based results into consistently strong whole-image-based results. Similar results have been found in other domains utilizing patch-based learning such as for cancer detection in whole slide images [28]. This is because even if patch-based results seem promising, they may not be clinically relevant. Machine learning practitioners must continue to address the issues of large data heterogeneity and small relative sizes of the regions of interest in the medical domain and report the appropriate user-relevant (clinician-relevant) metrics [41, 42, 43].

Acknowledgments and Disclosure of Funding

We acknowledge the support of the Mark Foundation for Cancer Research, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR), the Vector Institute for Artificial Intelligence, the University of Toronto, and The Hospital for Sick Children. We thank Dr. Babak Taati and Dr. Sanja Fidler for their valuable feedback.

References

[1] A. Ashwini and S. Murugan. Automatic skin tumour segmentation using prioritized patch based region – a novel comparative technique. *IETE Journal of Research*, 0(0):1–12, 2020.

[2] Soraya Gavazzi, Cornelis AT van den Berg, Mark HF Savenije, H Petra Kok, Peter de Boer, Lukas JA Stalpers, Jan JW Lagendijk, Hans Crezee, and Astrid LHMW van Lier. Deep learning-based reconstruction of in vivo pelvis conductivity with a 3d patch-based convolutional neural network trained on simulated mr data. *Magnetic resonance in medicine*, 84(5):2772–2787, 2020.

[3] L. Duran-Lopez, Juan P. Dominguez-Morales, D. Gutierrez-Galan, A. Rios-Navarro, A. Jimenez-Fernandez, S. Vicente-Diaz, and A. Linares-Barranco. Wide deep neural network model for patch aggregation in cnn-based prostate cancer detection systems. *Computers in Biology and Medicine*, 136:104743, 2021.

[4] Nikhil Naik, Ali Madani, Andre Esteva, Nitish Keskar, Michael Press, Dan Ruderman, David Agus, and Richard Socher. Deep learning-enabled breast cancer hormonal receptor status determination from base-level he stains. *Nature Communications*, 11, 11 2020.

[5] Mandy L. Ballinger, Ana Best, Phuong L. Mai, Payal P. Khincha, Jennifer T. Loud, June A. Peters, Maria Isabel Achatz, Rubens Chojniak, Alexandre Balieiro da Costa, Karina Miranda Santiago, Judy Garber, Allison F. O'Neill, Rosalind A. Eeles, D. Gareth Evans, Eveline Bleiker, Gabe S. Sonke, Marielle Ruijs, Claudette Loo, Joshua Schiffman, Anne Naumer, Wendy Kohlmann, Louise C. Strong, Jasmina Bojadzieva, David Malkin, Surya P. Rednam, Elena M. Stoffel, Erika Koeppe, Jeffrey N. Weitzel, Thomas P. Slavin, Bita Nehoray, Mark Robson, Michael Walsh, Lorenzo Manelli, Anita Villani, David M. Thomas, and Sharon A. Savage. Baseline Surveillance in Li-Fraumeni Syndrome Using Whole-Body Magnetic Resonance Imaging: A Meta-analysis. *JAMA Oncology*, 3(12):1634–1639, 12 2017.

[6] Hussna Elnoor Mohammed Abdalla and M. Y. Esmail. Brain tumor detection by using artificial neural network. In *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pages 1–6, 2018.

[7] Chirodip Lodh Choudhury, Chandrakanta Mahanty, Raghvendra Kumar, and Brojo Kishore Mishra. Brain tumor detection and classification using convolutional neural network and deep neural network. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–4, 2020.

[8] Tonmoy Hossain, Fairuz Shadmani Shishir, Mohsena Ashraf, MD Abdullah Al Nasim, and Faisal Muhammad Shah. Brain tumor detection using convolutional neural network. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6, 2019.

[9] Ahmet Çinar and Muhammed Yildirim. Detection of tumors on brain mri images using the hybrid convolutional neural network architecture. *Medical Hypotheses*, 139:109684, 2020.

[10] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[11] Alex Chang, Vinith M. Suriyakumar, Abhishek Moturu, Nipaporn Tewattanarat, Andrea Doria, and Anna Goldenberg. Using generative models for pediatric wbmri, 2020.

[12] Teppei Kanayama, Yusuke Kurose, Kiyohito Tanaka, Kento Aida, Shin'ichi Satoh, Masaru Kitsuregawa, and Tatsuya Harada. Gastric cancer detection from endoscopic images using synthesis by gan. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 530–538, Cham, 2019. Springer International Publishing.

[13] Shrinivas D Desai, Shantala Giraddi, Nitin Verma, Puneet Gupta, and Sharan Ramya. Breast cancer detection using gan for limited labeled dataset. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 34–39, 2020.

[14] Pooyan Sedigh, Rasoul Sadeghian, and Mehdi Tale Masouleh. Generating synthetic medical images by using gan to improve cnn performance in skin cancer classification. In *2019 7th International Conference on Robotics and Mechatronics (ICRoM)*, pages 497–502, 2019.

[15] Alex Chang, Vinith Suriyakumar, Abhishek Moturu, James Tu, Nipaporn Tewattanarat, Sayali Joshi, Andrea Doria, and Anna Goldenberg. 3d reasoning for unsupervised anomaly detection in pediatric wbmri, 2021.

[16] Aditya Khamparia, Deepak Gupta, Joel JPC Rodrigues, and Victor Hugo C de Albuquerque. Dcavn: Cervical cancer prediction and classification using deep convolutional and variational autoencoder network. *Multimedia Tools and Applications*, 80(20):30399–30415, 2021.

[17] Ning Xiao, Yan Qiang, Zijuan Zhao, Juanjuan Zhao, and Jianhong Lian. Tumour growth prediction of follow-up lung cancer via conditional recurrent variational autoencoder. *IET Image Processing*, 14(15):3975–3981, 2020.

[18] Feihong Li, Wei Huang, Mingyuan Luo, Peng Zhang, and Yufei Zha. A new vae-gan model to synthesize arterial spin labeling images from structural mri. *Displays*, 70:102079, 2021.

[19] Xiaofeng Liu, Fangxu Xing, Jerry L. Prince, Aaron Carass, Maureen Stone, Georges El Fakhri, and Jonghye Woo. Dual-cycle constrained bijective vae-gan for tagged-to-cine magnetic resonance image synthesis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1448–1452, 2021.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[21] Stefano Trebeschi, Joost JM van Griethuysen, Doenja MJ Lambregts, Max J Lahaye, Chintan Parmar, Frans CH Bakers, Nicky HGM Peters, Regina GH Beets-Tan, and Hugo JWL Aerts. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric mr. *Scientific reports*, 7(1):1–9, 2017.

[22] Kuo Men, Tao Zhang, Xinyuan Chen, Bo Chen, Yu Tang, Shulian Wang, Yexiong Li, and Jianrong Dai. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. *Physica Medica*, 50:13–19, 2018.

[23] Abhishek Moturu and Alex Chang. Creation of synthetic x-rays to train a neural network to detect lung cancer, 2018.

[24] Alex Chang and Abhishek Moturu. Detecting early stage lung cancer using a neural network trained with patches from synthetically generated x-rays, 2019.

[25] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547, 2016.

[26] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, oct 2020.

[27] Chensu Xie, Hassan Muhammad, Chad M. Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J. Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, 2020.

[28] Ozan Ciga, Tony Xu, Sharon Nofech-Mozes, Shawna Noy, Fang-I Lu, and Anne L Martel. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, 11(1):1–10, 2021.

[29] Satya P Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3d deep learning on medical images: a review. *Sensors*, 20(18):5097, 2020.

[30] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M. Zughaier, Muhammad Salman Khan, and Muhammad E.H. Chowdhury. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132:104319, 2021.

[31] Marc Moreno López, Joshua M Frederick, and Jonathan Ventura. Evaluation of mri denoising methods using unsupervised learning. *Frontiers in Artificial Intelligence*, 4:75, 2021.

[32] Malte Klingenberg, Didem Stark, Fabian Eitel, Kerstin Ritter, Alzheimer's Disease Neuroimaging Initiative, et al. Mri image registration considerably improves cnn-based disease classification. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 44–52. Springer, 2021.

[33] Yunfan Liu, Qi Li, Zhenan Sun, and Tieniu Tan. A 3 gan: an attribute-aware attentive generative adversarial network for face aging. *IEEE Transactions on Information Forensics and Security*, 16:2776–2790, 2021.

[34] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

[35] Ronald L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Inf. Process. Lett.*, 1:132–133, 1972.

[36] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022.

[37] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. Jun 2016.

[38] Rosana El Jurdi, Caroline Petitjean, Paul Honeine, and Veronika Cheplygina. A surprisingly effective perimeter-based loss for medical image segmentation.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[41] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

[42] David Bouget, André Pedersen, Asgeir S Jakola, Vasileios Kavouridis, Kyrre E Emblem, Roelant S Eijgelaar, Ivar Kommers, Hilko Ardon, Frederik Barkhof, Lorenzo Bello, et al. Preoperative brain tumor imaging: models and software for segmentation and standardized reporting. *Frontiers in neurology*, page 1500, 2022.

[43] Ying-Hwey Nai, Bernice W. Teo, Nadya L. Tan, Sophie O'Doherty, Mary C. Stephenson, Yee Liang Thian, Edmund Chiong, and Anthonin Reilhac. Comparison of metrics for the evaluation of medical segmentations using prostate mri dataset. *Computers in Biology and Medicine*, 134:104497, 2021.