Vision Transformer and Multiview Classification for Lesion Detection in 3D Cranial Ultrasound

Flora Estermann [1], *Valérie Kaftandjian* [2], *Philippe Guy* [2], *Philippe Quetin* [3], *Philippe Delachartre* [1]

[1] Univ Lyon, INSA Lyon, Université Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69621, Lyon, France

[2] Univ Lyon, INSA Lyon, Laboratoire Vibrations Acoustique (LVA), F-69621 Villeurbanne, France

[3] CH Avignon, France

ABSTRACT

With increasing advances in the field of medical brain imaging, we can now assess the presence of punctate white matter lesions (*PWML*) in the preterm infant. While some studies report a link between these lesions and adverse long-term outcomes, automatic detection of PWML through ultrasound (*US*) imaging could better assist doctors in diagnosis, at a lower cost than MRI. Many papers focus on MR biomedical image benchmark datasets, but few methods seem to tackle the detection of very small lesions in US images, because it is really challenging due to high class imbalance and low contrast. In this work, we propose a two-phase strategy: 1) Segmentation with a vision transformer to increase the number of detected lesions. 2) Multi-view classification of the lesions predicted in the output mask to reduce the number of false alarms and improve precision. We also compare 3 methods of preprocessing for input data. As a result, our method achieves better performances for PWML detection in US images compared to the best published models, with recall and precision reaching 82% and 60% respectively.

Index Terms— Deep Learning, Anomaly Detection, 3D Ultrasound, White Matter Injury, Vision Transformers.

1. INTRODUCTION

Punctate white matter lesions (*PWML*) usually appear during embryonic development in the central and periventricular regions of the brain. Their location is particularly important since it is linked to neurodevelopmental outcomes and can later cause cognitive and motor sequelae in early childhood [1]. Brain lesions are generally identified using MRI, but this procedure is expensive and not always accessible. As ultrasound is the paediatric routine, this modality could be of real interest for a broader screening of children with PWML and to assist medical diagnosis and treatment process.

However, the detection and segmentation of PWML on cranial ultrasound (cUS) is very challenging. First, despite a higher resolution than MRI, US images are difficult to analyse because of their low contrast, the presence of speckle and the high variability related to the data acquisition process. Second, PWML are very small (in our dataset, the median lesion volume is $1mm^3$), which leads to a significant imbalance in the data between lesion pixels and background pixels, making it difficult to train deep learning models.

Many papers present approaches using CNNs for tumors classification and segmentation, but few methods seem to tackle the detection of very small lesions, and even less in ultrasound. The problem of PWML detection in MRI was first adressed by Liu et al. [2], but despite the high contrast and low noise of MR images, the reported accuracy for the task remains low with a Dice under 0.60 and a recall at 0.65. In 2020, Erbacher et al. [3] started working on this task with cUS and introduced the Priority U-Net, with recall and precision reaching 0.53 and 0.50 respectively.

As stated above, the detection and segmentation of PWML on cUS is very difficult, not only because of the small size of lesions, but also considering the presence of numerous artifacts in cUS images that are very similar to true lesions in terms of gray levels, sometimes location or even shape and which tend to produce a large number of false positives and greatly reduce accuracy. In order to tackle this issue, we previously developed a 2 stage pipeline inspired by the work of Dakak et al. [4] on the detection of discontinuities in industrial CT volumes. A first step of oversegmentation was performed with the Priority U-Net trained on an expanded groundtruth, before applying a second step of multi-view classification with the 2.5D SC-Net [5] to reduce the number of false alarms. This approach achieved higher performance than existing methods, with recall and precision for the PWML detection reaching 0.72 and 0.56 respectively.

Motivated by the global context modeling capability of Vision Transformer [6], the TransUNet was proposed by Chen et al. [7] for medical image segmentation. As transformers have matched or even exceeded state-of-the-art in many applications, it was expected that integrating the TransUNet in our 2-step approach could also improve our results.

In this work, we demonstrate that performing segmentation with the TransUNet helps to increase the number of detected lesions (higher recall). In addition, we compare 3 methods of preprocessing for input data. Finally, the classification step that follows with the 2.5D SC-Net also helps to improve the precision of our results, while limiting the computational cost of adding more spatial context as well.

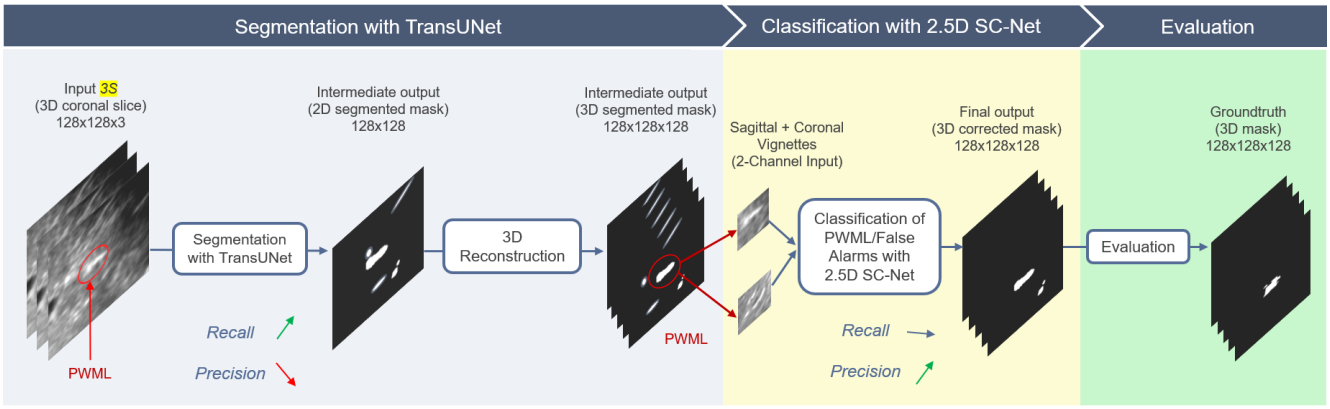

Fig. 1: Full pipeline : The Transunet takes 3 consecutive coronal slices as a 3-channel input and returns a segmented mask. The intermediate mask contains objects that may be PWML or false alarms. Hence step 2, where connected components are classified from sagittal and coronal views as a 2-channel input, to improve accuracy. In the end, the final mask is compared to the 3D groundtruth for evaluation.

2. METHODOLOGY

2.1. Segmentation with TransUNet

TransUNet [7] proved that Transformers could serve as powerful encoders for medical image segmentation tasks, with the combination of U-Net to enhance finer details by recovering localized spatial information. To add even more spatial context, 3 consecutive coronal slices from the US volume are given as a 3-channel input to the model 1.

Image sequentialization: We perform tokenization by reshaping the input image into a sequence of flattened 2D patches, where each patch is of size 16×16. A CNN is first used as a feature extractor to generate a feature map for the input image. In a second step, tokenized image patches are recovered from the CNN feature map and used as the input sequence for extracting global context with the transformer encoder.

Patch embedding: Patch embedding is applied to the patches extracted from the CNN feature map mentioned above. This process allows us to leverage the intermediate high-resolution CNN feature maps in the decoding path. Besides, it is found that the hybrid CNN-Transformer encoder performs better than simply using a pure Transformer as the encoder.

Cascaded upsampling (CUP): The decoder part of the network consists of multiple upsampling steps to decode the hidden feature for outputting the segmentation mask by combining the encoded features from the Transformer with the high-resolution CNN feature maps to enable precise localization. Besides, the CUP and the hybrid encoder actually form a U-shaped architecture which enables feature aggregation at different resolution levels via skip-connections. An intermediate 3D mask is then recovered by concatenating the 2D predictions along the coronal projection.

2.2. Classification of PWML & False alarms

The multi-view classifier 2.5D SC-Net [5] is trained on the joint fusion of the sagittal and coronal projections of the brain to differentiate true PWML from artifacts present in the brain, at a smaller scale but with more spatial context, while limiting the computational costs by using 2.5D instead of 3D volumes.

Patches of size $32 \times 32 \times 2$ centered on each connected components (CC) from the 3D intermediate mask are fed to the network as a 2-channel input (corresponding to each projection) to predict the class of the corresponding CC. During training, features are extracted through convolutional blocks for each projection separatly, then joint fusion is performed by computing the weighted average on the flatten output of the feature extraction part of the network. Note that the weights are learned automatically for each projection during training.

During the testing phase, once the 3D intermediate mask is obtained after the first step of segmentation with TransUNet, 2D patches are extracted around the regions of interest (thumbnails from the image, centered around the connected components of the predicted mask) from the sagittal and coronal projections of the brain, concatenated and sent to the 2.5D SC-Net to identify the true PWML. The intermediate mask is then corrected (i.e. the connected components predicted as false alarms are removed from the mask) 1.

3. EXPERIMENTS & RESULTS

3.1. Dataset

The 2D images are extracted from 54 reconstructed US brain volumes (including 29 with PWML) from 45 preterm babies whose mean age at birth was 31.6 ± 2.5 gestational weeks.

As the acquisition process of the ultrasound images is performed manually by the pediatrician along the anterior/posterior axis of the brain, the brain scan does not always results in the same number of dynamic sequences (DICOM).

In order to recover a complete volume, we completed this process by a reconstruction algorithm [8].

In total, the dataset without preprocessing contains 473 lesions. The smallest lesion barely reaches $0.03mm^3$, while the largest is more than $58mm^3$. The median lesion size is $1.08mm^3$, which is extremely tiny compared to what is observable in MRI. Besides, PWML have quite varied contrasts and do not really have specific shapes (punctate, ovoid or sometimes linear) 2. They are usually located in the center of hemispheres, near the lateral ventricles.

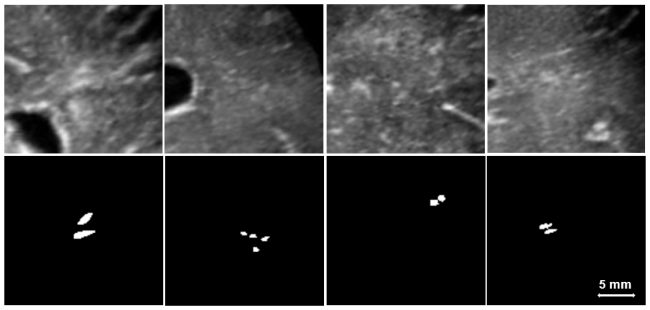

Fig. 2: PWML examples from the Brain US Dataset (Top row : US images. Bottom row : groundtruth masks). PWML have varied contrasts and shapes, and are often difficult to distinguish from peripheral vessels or arteries in cross-section.

3.2. Data Preprocessing

A first preprocessing phase consists of extracting a sub-volume of size $128 \times 128 \times 128$ in the top-right hemisphere, periventricular region of the brain for each patient.

In order to reduce class imbalance, a first filtering is performed on the size of the lesions for each volume to limit the number of lesions that are too small and to make the problem less complex. As a result, only 90% of the lesional volume is kept for each patient, which allows us to get rid of the tiniest lesions, that are usually not even visible in the MRI.

Additionally, several methods of data preprocessing were explored separately in order to obtain the optimal input for the TransUNet :

Duplicated grayscale (2D): The input to the network is a 2D grayscale coronal slice duplicated 3 times and given as a 3-channel input of dimensions $3 \times 128 \times 128$ to the segmentation network.

Expanded groundtruth (HF): The PWML in the groundtruth are artificially expanded by aggregating the foreground pixels within a 5-slices sliding window along the coronal projection of the brain. This results in a mask with a higher percentage of foreground pixels which is expected to cause additional loss and helps to make training more effective.

3 consecutive grayscale (3S): The input to the TransUNet is the concatenation of 3 consecutive 2D slices extracted from the volume along the coronal projection of the brain, given as a 3-channel input of dimensions $3 \times 128 \times 128$ to the model. The idea was to use these 3 channels to give the model more spatial information. Our intuition was that thickening the lesions seen by the model might help it to over-segment the groundtruth, and thus better detect small lesions.

Horizontal flipping and rotation are randomly applied to the chosen input with a probability of 0.5 for data augmentation.

3.3. Experimental Setup

The proposed pipeline [1] was implemented in Python 3.8 with PyTorch (TransUNet) and TensorFlow (2.5D SC-Net) backend. All the models were trained and tested with GPU. For each model, we performed a 10-fold cross-validation and computed the median of scores.

The TransUNet was trained for about 40 epochs with the Dice Loss, whereas the 2.5D SC-Net was trained for approximately 20 epochs using the Weighted-Binary Cross-Entropy Loss. The batch size is 4 for the segmentation and 32 for the classification. The initial learning rate was fixed at 10e-3 with the Adam optimizer and automatically decreased by a factor 0.1 when validation loss did not improved after 10 epochs.

3.4. Results

Table 1: Final results of the proposed approach (TransUNet (3S) + 2.5D SC-Net) compared to state-of-the-art. All these results are the medians of 10-folds cross-validation (± median absolute deviation).

Model	Recall	Precision	Dice *(TP)*
U-Net [9]	55.83 (± 8)	58.59 (± 10)	47.21 (± 12)
Priority U-Net [3]	71.63 (± 12)	48.96 (± 11)	59.03 (± 6)
Priority U-Net (HF) + 2.5D SC-Net [5]	79.02 (± 8)	56.35 (± 10)	58.89 (± 9)
TransUNet (2D) [7]	80.32 (± 11)	50.65 (± 9)	62.22 (± 8)
TransUNet (HF)	78.13 (± 11)	56.25 (± 12)	58.10 (± 9)
TransUNet (3S)	**82.31** (± 8)	51.56 (± 10)	62.20 (± 12)
TransUNet (3S) + 2.5D SC-Net	**82.19** (± 9)	**60.00** (± 9)	**66.63** (± 4)

To quantitatively assess the quality of the PWML detection produced by the target pipeline, we employed 3 criteria to evaluate each model : the Recall and the Precision for the detection task but also the Dice on true positives *(TP)* to get an overview of the segmentation ability of the model. For each of these metrics, the value closer to 1 the better. The quantitative results are shown in Table 1. Note that the high

[1]The source code is available here: `https://github.com/FloWPs/PWML_Automatic_Detection_IUS_2023`

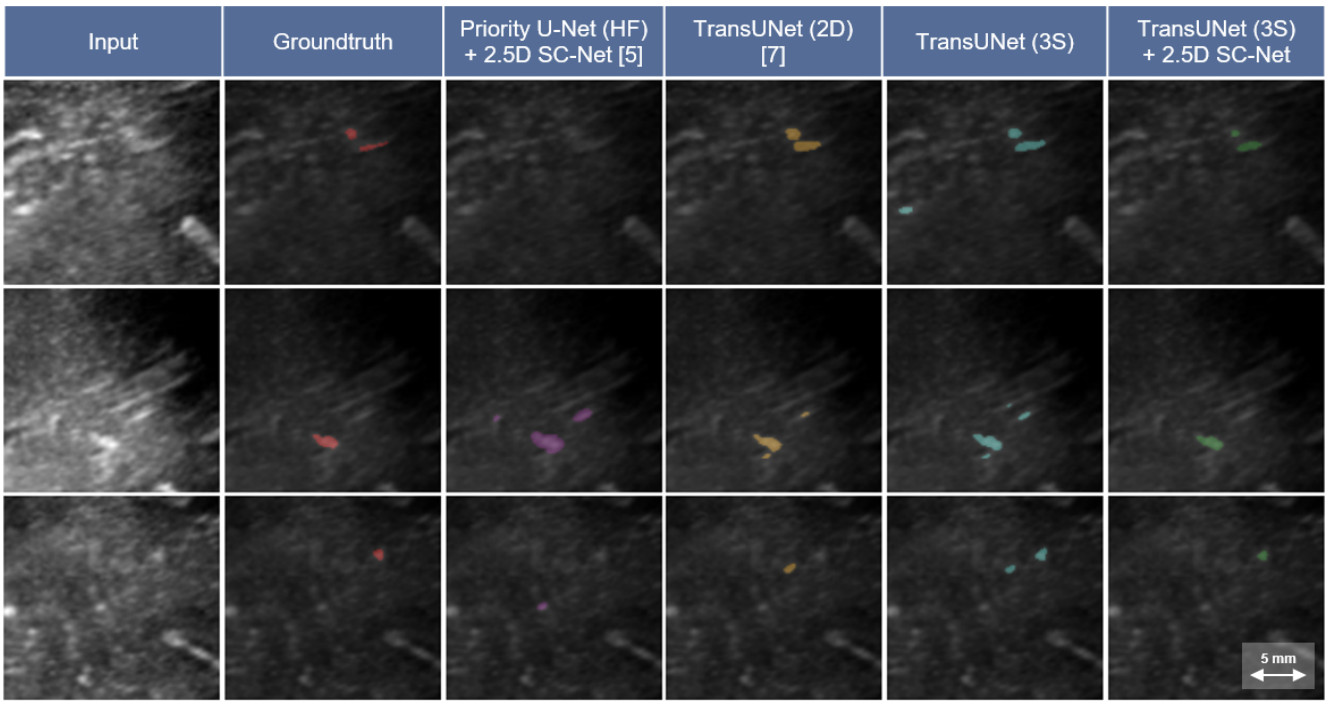

Fig. 3: Visual examples of the PWML detection with our method (TransUNet 3S + 2.5D SC-Net) compared to state-of-the-art approaches. The input is the US image given to the network, the corresponding groundtruth is in the second column and the last 4 columns show the comparison of predictions from the different models.

variability in results may be explained by the limited number of patients available for each validation fold.

In our study, we observe that expanding the lesions to train the TransUNet (HF) does not necessarily improve results. On the other hand, giving 3 consecutive slices as a 3-channel input (3S) to the model slightly improves recall and precision (+2% and +1% respectively). Besides, as shown in the paper [5], we demonstrate that applying a second step of multi-view classification with the 2.5D SC-Net after the segmentation significantly helps to improve the precision (+9%) and the dice (+4%) with nearly no impact on the recall.

This can also be illustrated in Fig. 3, where we find that our full pipeline (last column) tend to include fewer false positives after classification, while still detecting PWML better than most other approaches.

As a result, our model achieves better performances for PWML detection in US images compared to other methods.

4. DISCUSSION AND CONCLUSION

Detecting PWML in US is challenging due to high class imbalance and low contrast imaging. By giving 3 consecutive slices to the TransUNet and thickening the input, the over segmentation allows a better detection of PWML. During the classification step, giving more spatial context as a 2 channel input and implementing joint fusion in the multiview classifier 2.5D SC-Net helps to reduce the number of false alarms and also to improve the segmentation performance. At the end of the proposed pipeline, we reach a higher recall and precision (82% and 60% respectively) than those obtained with other state-of-the-art techniques.

While most people have conducted this task on MR images, this work highlights once again the possibility of detecting brain lesions through ultrasound imaging.

However, we are aware that the limited number of patients available for each validation fold induces high variability in the results. Hence future works will focus on enriching the database. Besides, we also plan to integrate attention mechanisms into the multi-view classifier as well.

5. ACKNOWLEDGMENTS

This work was supported by the LABEX CELYA operated by the French National Research Agency (ANR) and the University of Lyon, within INSA Lyon.

6. COMPLIANCE WITH ETHICAL STANDARDS

The data from human subjects used in this work were obtained and treated in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committees of the institutions involved in creating the PWML database, from which these data were accessed. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Nguyen A.L.A., Ding Y., Suffren S., Londono I., Luck D., and Lodygensky G.A., "The brain's kryptonite: Overview of punctate white matter lesions in neonates," *International Journal of Developmental Neuroscience*, vol. 77, no. 1, pp. 77–88, October 2019.

[2] Liu Y., Li J., Wang M., Jiao Z., Yang J., and Li X., "Trident segmentation cnn: A spatiotemporal transformation cnn for punctuate white matter lesions segmentation in preterm neonates," *Med Biol Eng Comput*, 2019.

[3] Erbacher P., Lartizien C., Martin M., Foletto Pimenta P., Quetin P., and Delachartre P., "Priority u-net: Detection of punctuate white matter lesions in preterm neonate in 3d cranial ultrasonography," *Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL)*, 2020.

[4] Dakak A.R., Kaftandjian V., and Duvauchelle P.and Bouvet P., "Deep learning-based defect detection in industrial ct volumes of castings," *Insight Non-Destructive Testing and Condition Monitoring*, vol. 64, no. 11, November 2022.

[5] Estermann F., Kaftandjian V., Guy P.and Quetin P., and Delachartre P., "Pwml detection in 3d cranial ultrasound volumes using over-segmentation and multimodal classification with deep learning," *IEEE International Symposium on Biomedical Imaging*, 2023.

[6] Dosovitskiy A., Beyer A., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., and Houlsby N., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[7] Chen J., Lu Y., Yu Q., Luo X., Adeli E., Wang Y., Lu L., Yuille A., and Zhou Y., "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[8] Martin M., Sciolla B., Sdika M., Wang X., Quetin P., and Delachartre P., "Automatic segmentation of the cerebral ventricle in neonates using deep learning with 3d reconstructed freehand ultrasound imaging," *IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, 2018.

[9] Ronneberger O., Fischer P., and Brox T., "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, pp. 234–241, 2015.