Transformer-based Model for Oral Epithelial Dysplasia Segmentation

Shephard, Adam J; Mahmood, Hanya; Raza, Shan E Ahmed; Araujo, Anna Luiza Damaceno; Santos-Silva, Alan Roger; Lopes, Marcio Ajudarte; Vargas, Pablo Agustin; McCombe, Kris; Craig, Stephanie; James, Jacqueline; Brooks, Jill; Nankivell, Paul; Mehanna, Hisham; Khurram, Syed Ali; Rajpoot, Nasir M

DOI:
10.48550/arXiv.2311.05452

TRANSFORMER-BASED MODEL FOR ORAL EPITHELIAL DYSPLASIA SEGMENTATION

Adam J Shephard[1], *Hanya Mahmood*[2], *Shan E Ahmed Raza*[1], *Anna Luiza Damaceno Araujo*[3], *Alan Roger Santos-Silva*[3], *Marcio Ajudarte Lopes*[3], *Pablo Agustin Vargas*[3], *Kris McCombe*[4], *Stephanie Craig*[4], *Jacqueline James*[4], *Jill Brooks*[5], *Paul Nankivell*[5], *Hisham Mehanna*[5], *Syed Ali Khurram*[2*], *Nasir M Rajpoot*[1*]

[1] Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, UK
[2] School of Clinical Dentistry, University of Sheffield, UK
[3] Oral Diagnosis Department, Piracicaba Dental School University of Campinas, São Paulo, Brazil
[4] Precision Medicine Centre, Queen's University Belfast, UK
[5] Institute of Head and Neck Studies and Education, University of Birmingham, UK
[*] Joint co-senior authorship

ABSTRACT

Oral epithelial dysplasia (OED) is a premalignant histopathological diagnosis given to lesions of the oral cavity. OED grading is subject to large inter/intra-rater variability, resulting in the under/over-treatment of patients. We developed a new Transformer-based pipeline to improve detection and segmentation of OED in haematoxylin and eosin (H&E) stained whole slide images (WSIs). Our model was trained on OED cases (n = 260) and controls (n = 105) collected using three different scanners, and validated on test data from three external centres in the United Kingdom and Brazil (n = 78). Our internal experiments yield a mean F1-score of 0.81 for OED segmentation, which reduced slightly to 0.71 on external testing, showing good generalisability, and gaining state-of-the-art results. This is the first externally validated study to use Transformers for segmentation in precancerous histology images. Our publicly available model shows great promise to be the first step of a fully-integrated pipeline, allowing earlier and more efficient OED diagnosis, ultimately benefiting patient outcomes.

Index Terms— Oral Epithelial Dysplasia, Segmentation, Tranformer, Computational Pathology, Histopathology

1. INTRODUCTION

Oral epithelial dysplasia (OED) presents a significant challenge in the realm of head & neck pathology, where accurate diagnosis and early detection are paramount for effective intervention and the prevention of malignant progression [1]. OED is a premalignant histopathological diagnosis encompassing various lesions of the oral mucosa, typically manifesting as white (leukoplakia), red (erythroplakia) or mixed (red-white) lesions [1]. Accurate diagnosis and early detection of OED are crucial for effective intervention and prevention of malignant progression. However, the current manual assessment of H&E-stained sections of oral tissue slides, the gold standard in OED diagnosis, suffers from low throughput and susceptibility to intra-/inter-observer variability [1, 2].

To address these challenges and enhance the diagnosis and management of OED, there is a growing interest in leveraging advanced technologies, particularly deep learning, which has seen extensive use in medical image analysis over the last decade [3, 4]. Concurrently, Transformers have captured widespread attention in recent years due to their successful application in various domains, including natural language processing and computer vision tasks, such as classification [5]. A typical Transformer encoder comprises three fundamental components: a multi-head self-attention (MSA) layer, a multi-layer perceptron (MLP), and layer normalisation (LN). The inclusion of the MSA layer is particularly noteworthy as it empowers Transformers to capture long-range dependencies, rendering them a promising choice for semantic segmentation in the context of medical images [6, 7]. While Transformers have demonstrated their potential to mitigate some of the constraints associated with convolutional neural networks (CNNs), their utilization in histological applications has been primarily limited to classification tasks, with semantic segmentation left relatively unexplored. This raises the question of whether Transformers can be harnessed for segmentation of histological images.

In this study, we apply a Transformer-based model to a comprehensive OED dataset for dysplasia segmentation, setting a new standard in the field. Our model is built on the Trans-UNet architecture [6], and is specifically designed for segmenting dysplastic regions in H&E-stained whole slide images (WSIs) of oral tissue. We believe that the application of cutting-edge, state-of-the-art (SOTA) deep learning techniques, such as Transformer-based architectures, holds the potential to significantly improve the accuracy and effi-

Table 1. Internal testing results with different loss functions and patch sizes/resolutions.

Loss	Patch Size	Res. (mpp)	OED cases			Controls
			F1	Recall	Prec.	Spec.
Dice + CE	256	1.0	0.794	0.824	0.767	0.998
Dice + CE	512	0.5	0.781	0.792	0.771	0.999
Dice + CE	**512**	**1.0**	**0.807**	0.844	0.773	0.997
Dice	512	1.0	0.795	**0.852**	0.746	0.996
Jaccard	512	1.0	0.000	0.000	0.000	**1.000**
CE	512	1.0	0.805	0.834	**0.778**	0.998
Jaccard + CE	512	1.0	0.784	0.828	0.744	0.996

ciency of OED diagnosis. We rigorously evaluate the performance of our model by comparing it to other SOTA methods, and demonstrate its robustness and generalisability by extending our evaluation to include cases from three external and international centres: Birmingham (UK), Belfast (UK) and São Paulo (Brazil). We have open-sourced our model inference pipeline to facilitate broader research and application (https://github.com/adamshephard/oed_inference).

2. METHOD

2.1. Study Data

2.1.1. Training Data

The training dataset comprised a retrospective sample of histology tissue sections collected (dating 2008 to 2016) from the Oral and Maxillofacial Pathology archive at the School of Clinical Dentistry, University of Sheffield, UK. New tissue sections of the selected cases were cut (4 μm thickness) from formalin fixed paraffin embedded (FFPE) blocks and stained with H&E. The dataset comprised 260 slides with a histological diagnosis of OED, and 105 non-dysplastic (control) slides. Slides were scanned at 40$\times$ objective power with either a NanoZoomer S360 (Hamamatsu Photonics, Japan; 0.2258 mpp), an Aperio CS2 (Leica Biosystems, Germany; 0.2520 mpp), or a Pannoramic 1000 (3DHISTECH Ltd., Hungary; 0.2426 mpp) slide scanner to obtain digital WSIs. Exhaustive delineation of ROIs representing dysplastic epithelium in OED slides, and normal epithelium in controls slides, was performed using QuPath [8].

2.1.2. External Testing Data

For the external testing of the models generated in this study, we recruited OED cases from three external centres: (i) Queen's University Belfast, UK; (ii) Institute of Head and Neck Studies and Education, Birmingham, UK; and (iii) Piracicaba Dental School, Brazil. 30 OED cases were collected from Belfast, 30 from Birmingham and 18 from Brazil. The Birmingham and Belfast slides were scanned at 40$\times$ objective power using a Pannoramic 250 (P250, 3DHISTECH Ltd., Hungary; 0.1394 mpp), and an Aperio AT2 (Leica Biosystems, Germany; 0.2529 mpp) scanner, respectively. The Brazil cases were scanned at 20$\times$ objective power, by an Aperio CS (Leica Biosystems, Germany; 0.4928 mpp) scanner. Owing to the limited size of the these datasets we combined them into a single multi-institutional test set consisting of 78 OED cases. Exhaustive delineation of dysplastic ROIs in the epithelium in all cases was performed.

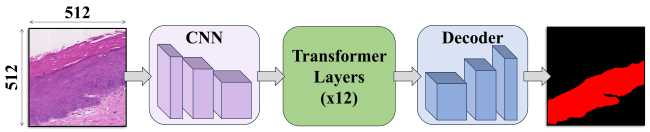

Fig. 1. Network architecture of the TransUNet model.

2.2. Network Architecture and Implementation

We present a new model for OED segmentation, based on the TransUNet [6] architecture (see Fig. 1). This is a hybrid model, that uses a CNN (ResNet50 [9]) as a feature extractor. 1×1 patches are then extracted from the feature maps and used for patch embedding for the Transformer layers. Finally, a cascaded upsampler is used as a decoder. This allows feature aggregation through skip-connections, thus leveraging the high-resolution CNN feature maps in the decoding path.

Our model takes an input RGB image of size 512×512 (at 1.0 micron per pixel, mpp, resolution) and outputs a dysplasia segmentation map. For post-processing, we performed morphological closing/opening, and removal of small objects and holes. We first tested the proposed model over varying patch sizes, resolutions, and loss functions. To aid our model in generalising to unseen domains, we tested its performance based on various domain generalisation (DG) techniques. Methods we employed included: weighted sampling (WS), stain augmentation (SA), and domain adversarial training (DA) [10]. Following this, we compared our model against other state-of-the-art deep learning models for semantic segmentation, including Swin-UNet [7], U-Net [11] (ResNet-50 [9] backbone), Efficient-UNet [12] (EfficientNet-B7 backbone), DeepLabV3+ [13] (ResNet-101 backbone), and HoVer-Net+ ([14]; segmentation decoder alone).

All of these models were trained based on their default parameters, and pretrained on ImageNet.

We trained all models in two phases. We trained the decoders for 20 epochs first, before training the entire network for 30 epochs second. The Adam optimizer was used with a learning rate that decayed initially from 10^{-4} to 10^{-5} after 10 epochs, in both phases. We applied the following random data augmentations: flip, rotation, Gaussian blur, median blur, and colour perturbation. We additionally tested the effect of stain augmentation using the TIAToolbox [15] implementation of the Macenko method [16]. This has been shown previously to help counter scanner-induced domain-shift [17, 18, 19].

For internal testing, we split the dataset with a 80/20 split controlled for both scanner and OED grade. This resulted in 206 OED and 75 control slides in the training set, and 54 OED and 21 control slides in the testing set. A higher proportion of controls were kept in the test set to ensure a high specificity of OED segmentation in controls. An equal number of cases and controls were used from each scanner in the test set. We tessellated our WSIs and masks into smaller patches of size 512×512 (overlap of 184) pixels at $10\times$ magnification (1.0 mpp), resulting in a total of 11,756 normal patches and 19,063 OED patches for model training/validation on the discovery cohort. For model testing, we report our evaluation metrics at the ROI level. Typically, each case/control had only one complete tissue section annotated. However, since some of the WSIs contained multiple tissue sections with annotations, this amounted to 66 OED ROIs and 23 control ROIs for testing. Each ROI encapsulated a whole tissue section.

For external validation, we trained our models based on the Sheffield data, and tested on the 78 external OED cases. This resulted in a total of 6,341 OED patches for model validation. Since some of these WSIs contained multiple tissue sections with annotations, this totalled 87 OED ROIs. The external data only comprises OED cases (and no controls).

2.3. Evaluation Metrics

For OED cases, we report an F1-score, recall and precision, aggregated over all ROIs. For controls, we provide the model specificity, since a single false positive pixel, would result in F1, recall, and precision values of 0; thus not giving an accurate representation of the model performance.

3. EXPERIMENTS AND RESULTS

We first tested the performance of our model over differing patch sizes (and resolutions), and loss functions; where we found a patch size of 512×512 at 1.0 mpp, with a combined Dice and cross-entropy loss function to be best (see Table 1). Next, we tested the proposed model when comparing the incorporation of various domain generalisation techniques (see Table 2). These techniques yielded no improvement in performance on internal testing, with domain adversarial training hindering performance. Stain augmentation improved specificity on controls, with a slight reduction in F1-score. We suggest that these techniques were not beneficial on internal testing as slides from all three scanners were present in both the training and testing set. Instead, techniques such as stain augmentation may be more beneficial for external testing.

We compared our model to other state-of-the-art methods in Table 3. Here, we see the superiority of the proposed model (F1 = 0.81) when compared to all other models. DeepLabV3+ was the closest performing model (F1 = 0.80), with U-Net being worst (F1 = 0.78). We additionally provide a dysplasia heatmap for a severe OED case (see Fig. 2), showing our model's accurate segmentation. The proposed model generalised well on external testing, gaining an F1-score of 0.71, and a high recall (see Table 4). Stain augmentation did not appear to improve the model F1-score; however, it did make the model more precise. We provide the comparative model results in Table 4, showing our proposed model to be best.

Table 2. Internal testing on the OED cases and controls, whilst testing domain generalisation techniques.

DG Method	OED cases			Controls
	F1	Recall	Prec.	Spec.
WS	0.798	0.839	0.760	0.998
SA [16]	0.805	**0.858**	0.758	0.997
DA [10]	0.682	0.723	0.644	0.984
WS, SA	0.802	0.851	0.758	0.997
WS, DA	0.700	0.749	0.657	0.991
SA, DA	0.735	0.774	0.701	0.992
WS, SA, DA	0.699	0.725	0.655	0.988
Proposed	**0.807**	0.845	**0.773**	**0.998**

Table 3. Comparative experiments for internal testing.

Model	OED cases			Controls
	F1	Recall	Prec.	Spec.
U-Net [11]	0.775	0.796	0.755	0.996
HoVer-Net+ [14]	0.789	0.827	0.754	0.996
DeepLabV3+ [13]	0.802	0.817	**0.788**	0.998
Efficient-UNet [12]	0.790	0.834	0.751	0.998
Swin-UNet [7]	0.795	0.845	0.750	0.997
Proposed	**0.807**	**0.845**	0.773	**0.998**

Table 4. Comparative experiments for external testing.

Model	F1	Recall	Prec.
U-Net [11]	0.685	0.694	0.676
HoVer-Net+ [14]	0.668	0.719	0.623
DeepLabV3+ [13]	0.704	0.704	**0.705**
Efficient-UNet [12]	0.700	**0.777**	0.638
Swin-UNet [7]	0.680	0.728	0.638
Proposed	**0.708**	0.764	0.660
Proposed (SA)	**0.708**	0.744	0.676

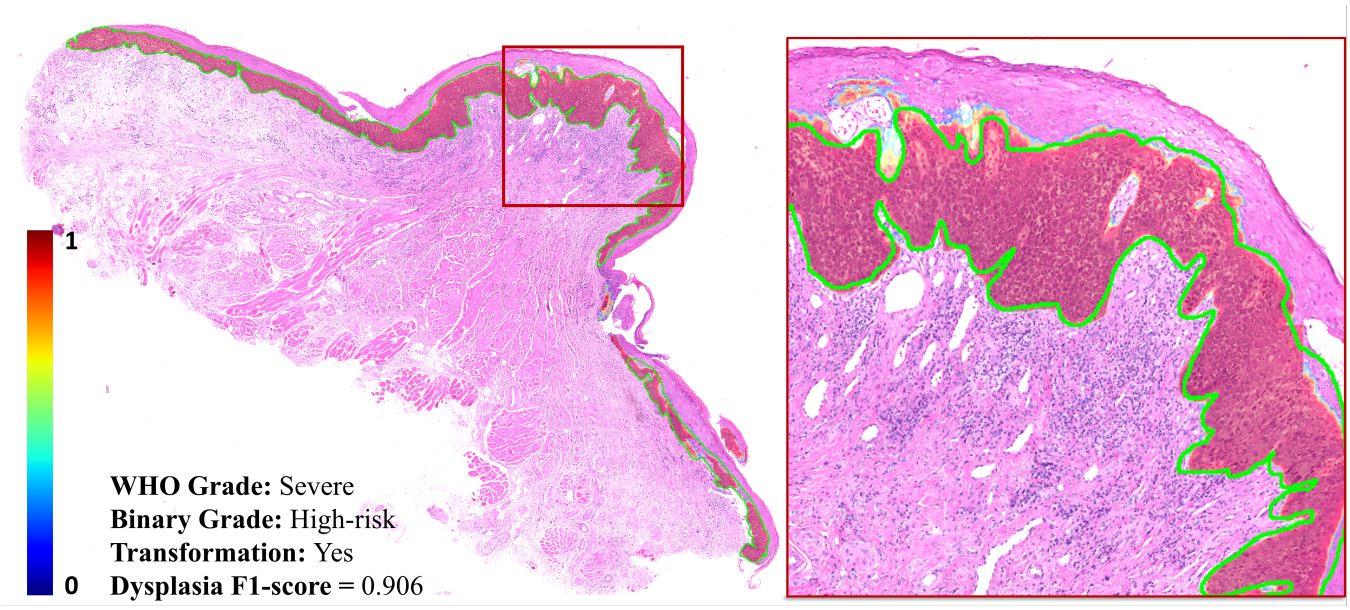

Fig. 2. Dysplasia segmentation heatmap for a case graded as Severe (WHO Grade) and High-risk (Binary Grade), that transformed to cancer. The green line is the ground truth dysplasia segmentation by the pathologist. The red box shows a zoom in of one the most dysplastic areas.

4. DISCUSSION AND CONCLUSION

In this study, we presented a Transformer-based model tailored for OED segmentation, and introduced the most extensive and diverse OED dataset to date. This dataset comprises 338 OED slides from four global centres, scanned using six different digital slide scanners, along with 105 control slides. Our study represents the first successful application of Transformer-based architectures for semantic segmentation in head & neck histology images. Our model's architecture, featuring the Transformer's self-attention mechanism, enables it to capture long-range dependencies, making it well-suited for segmentation in complex medical images. Our model achieved remarkable performance, consistently outperforming other SOTA deep learning models. This underlines the technical prowess of Transformer-based architectures in tackling challenging medical image segmentation tasks.

We found one other study to perform OED segmentation [20]. This study focussed on moderate/severe OED cases, where dysplasia is more pronounced, and achieved an F1-score of 0.64 for segmentation at the patch-level. In comparison, all of our metrics are provided at the ROI-level, a harder task due to containing a higher variation of tissue type. Even so, we have clearly surpassed this performance on both internal (F1 = 0.81) and external testing (F1 = 0.71).

A key technical achievement of our model is its robustness across diverse data sources. By training the model on slides from various scanners, we addressed a fundamental challenge in medical image analysis: domain shift [21]. This ensured our model maintained its performance, even in the presence of variations introduced by different scanners/sites. Its ability to generalise well across external datasets, is a crucial indicator of its robustness and applicability in diverse clinical settings.

As we move forward, the integration of the proposed model into clinical practice holds promise for enhancing the efficiency and reliability of OED diagnosis. Future research should focus on the seamless integration of this model into the assessment of individual slides for OED diagnosis and grading, enabling swifter and more objective treatment. Finally, the external validation of our models across multiple centres/scanners is a notable strength of this study. Future research could explore the application of the proposed model in even more diverse clinical settings and expand its utility to other histopathological tasks beyond OED. We suggest testing the method on other precancerous squamous lesions, such as laryngeal and cervical dysplasia, or even other types of dysplasia such as ductal carcinoma *in situ*.

In conclusion, our research represents a substantial advancement in head & neck pathology by providing a powerful publicly available model for OED segmentation, powered by a Transformer-based architecture. This technology demonstrates the transformative potential of computational pathology in improving the diagnosis and management of OED. As we address challenges and refine the model, deep learning is poised to play a vital role in enhancing the diagnosis of head & neck precancerous lesions in the future. Finally, this works serves as a benchmark for future research into the use of Transformers for segmentation in histopathology images.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Ethical approval was granted by the NHS Health Research Authority West Midlands (18/WM/0335)

6. ACKNOWLEDGEMENTS

This work was supported by a Cancer Research UK Early Detection Project Grant (C63489/A29674), and a National Institute for Health Research grant (NIHR300904).

7. REFERENCES

[1] Paul M. Speight, Syed Ali Khurram, and Omar Kujan, "Oral potentially malignant disorders: risk of progression to malignancy," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 125, no. 6, pp. 612–627, 2018.

[2] Omar Kujan et al., "Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: An attempt to understand the sources of variation," *Oral Oncology*, vol. 43, no. 3, pp. 224–231, 2007.

[3] Geert Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[4] Anant Madabhushi and George Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis*, vol. 33, pp. 170–175, 2016.

[5] Kelei He et al., "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[6] Jieneng Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv*, pp. 1–13, 2021.

[7] Hu Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[8] Peter Bankhead et al., "QuPath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, no. 1, pp. 1–7, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[10] Yaroslav Ganin et al., "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[12] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar, "Eff-UNet: A novel architecture for semantic segmentation in unstructured environment," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2020-June, pp. 1473–1481, 2020.

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[14] Adam J Shephard et al., "Simultaneous Nuclear Instance and Layer Segmentation in Oral Epithelial Dysplasia," *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[15] Johnathan Pocock et al., "TIAToolbox as an end-to-end library for advanced tissue image analytics," *Communications Medicine*, vol. 2, no. 1, pp. 120, 2022.

[16] Marc Macenko et al., "A method for normalizing histology slides for quantitative analysis," *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, pp. 1107–1110, 2009.

[17] Marc Aubreville et al., "Mitosis domain generalization in histopathology images — The MIDOG challenge," *Medical Image Analysis*, vol. 84, pp. 102699, feb 2023.

[18] Mostafa Jahanifar et al., "Stain-Robust Mitotic Figure Detection for the Mitosis Domain Generalization Challenge," *arXiv*, pp. 3–5, 2021.

[19] Mostafa Jahanifar et al., "Stain-Robust Mitotic Figure Detection for MIDOG 2022 Challenge," *arXiv*, 2022.

[20] Yingci Liu, Elizabeth Bilodeau, Brian Pollack, and Kayhan Batmanghelich, "Automated detection of premalignant oral lesions on whole slide images using convolutional neural networks," *Oral Oncology*, vol. 134, pp. 106109, 2022.

[21] Mostafa Jahanifar et al., "Domain generalization in computational pathology: Survey and guidelines," *arXiv*, 2023.