Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Slater-determinant exclusion gating does not improve polyp segmentation: a controlled evaluation with a prespecified analysis plan

Amine BOUGARNE

amine.bougarne@um5r.ac.ma

Mohammed V University

Article

Keywords: colonoscopy, polyp segmentation, Slater determinant, feature redundancy, determinantal point processes, negative results, reproducibility, Kvasir-SEG

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11381150/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Slater-determinant exclusion gating does not improve polyp segmentation: a controlled evaluation with a prespecified analysis plan

Amine BOUGARNE[a,*]

[a] *Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco.*

*Corresponding author. E-mail: amine.bougarne@um5r.ac.ma

ORCID: 0009-0000-6896-6926 (A. BOUGARNE)

Abstract

Polyp segmentation networks learn spatially redundant features and often misfire on specular highlights. Determinants give an exact test of redundancy, but layers built on them are seldom compared with controls that isolate the determinant. We built a Slater exclusion gate that scores each feature token by its Schur-complement distance to the span of its neighbours, a score that is zero exactly when the token is linearly dependent on them. On Kvasir-SEG (1,000 images; grouped five-fold cross-validation repeated with three seeds) we compared it with a U-Net, a parameter-matched gate without token interaction and a pairwise-exclusion gate, under an analysis plan fixed before training, and we tested every model without retraining on four external datasets. The gate did not improve per-image Dice over its matched control (0.9006 vs 0.9007; difference −0.0001, 95% CI −0.0031 to 0.0029), and it did not reduce false positives inside highlights. Gate diagnostics show why: it attenuated almost every deep token (mean gate 0.08 at stride 32). Adding the same quantity to the loss as a log-determinant penalty lowered inter-token redundancy by 14% with Dice comparable to the U-Net (0.9054 vs 0.9038). Code, splits and verified reimplementations of PraNet and Polyp-PVT are released.

Keywords: colonoscopy; polyp segmentation; Slater determinant; feature redundancy; determinantal point processes; negative results; reproducibility; Kvasir-SEG

Introduction

Colorectal cancer caused more than 1.9 million new cases and about 935,000 deaths worldwide in 2020[1]. Colonoscopy prevents part of this burden by finding and removing adenomas, and its protective effect depends on how many of them the endoscopist sees: in a cohort of 314,872 examinations, each 1% increase in the adenoma detection rate was associated with a 3% lower risk of interval cancer[2]. Adenomas are still missed. A meta-

analysis of tandem colonoscopy studies put the adenoma miss rate at about one in four[3]. Computer assistance that delineates polyps in real time is one answer, and pixel-level segmentation is its most demanding form.

Public datasets made the problem measurable. Kvasir-SEG provides 1,000 polyp images with expert masks[4], and CVC-ClinicDB[5], CVC-ColonDB[6], ETIS-LaribPolypDB[7] and the EndoScene test set CVC-300[8] come from other centres and devices. Since PraNet[9], most papers train on 900 Kvasir-SEG and 550 CVC-ClinicDB images and report a mean Dice averaged over 256 binarisation thresholds on five test sets, which makes the field easy to compare but also easy to overfit to one protocol.

Architectures have moved quickly on this benchmark. U-Net[10] and UNet++[11] set the first baselines, and ResUNet++ added residual and attention blocks[12,13]. SFA modelled area and boundary jointly[14]. PraNet introduced parallel reverse attention on a Res2Net backbone[9,15], and later models refined the decoder with multi-scale subtraction (MSNet[16]), shallow attention and colour exchange (SANet[17]), uncertainty-aware context (UACANet[18]), axial reverse attention (CaraNet[19]) or a lightweight HarDNet encoder[20]. Transformers then entered the field with TransFuse[21] and Polyp-PVT[22], which pairs a PVTv2 encoder[23] with cascaded fusion and graph-based similarity aggregation. Under the PraNet protocol, the reported mean Dice on Kvasir-SEG rose from 0.818 for U-Net to 0.917 for Polyp-PVT (Table 1). Reviews of image and video polyp segmentation describe this progression in detail[24,25], and nnU-Net reminds us that a well-configured plain baseline is often harder to beat than a new module suggests[26].

Table 1. Published mean Dice under the PraNet protocol (training on 900 Kvasir-SEG and 550 CVC-ClinicDB images; 352 × 352 input; Dice averaged over 256 thresholds). Values are taken from the paper that introduced each method.

Method	Year	Kvasir	ClinicDB	ColonDB	ETIS	CVC-300
U-Net[10]	2015	0.818	0.823	0.512	0.398	0.710
UNet++[11]	2018	0.821	0.794	0.483	0.401	0.707
SFA[14]	2019	0.723	0.700	0.469	0.297	0.467
PraNet[9]	2020	0.898	0.899	0.709	0.628	0.871
SANet[17]	2021	0.904	0.916	0.753	0.750	0.888
Polyp-PVT[22]	2023	0.917	0.937	0.808	0.787	0.900

U-Net, UNet++ and SFA were evaluated by the PraNet authors. Later tables report 0.712 for PraNet on CVC-ColonDB (0.709 in the original paper) and, in one secondary table of Polyp-PVT, 0.456 for UNet++ on CVC-ColonDB (0.483 in the original); see Supplementary Table S3. Year refers to the original method.

Two weaknesses recur in this literature. The first is artefacts. Specular highlights, bubbles and mucus share the brightness and texture of polyp surfaces, and highlights alone have motivated

dedicated detection and inpainting methods[27]. The second is redundancy. Neighbouring positions of a deep feature map carry nearly the same information, and several lines of work treat this as a problem to be penalised: decorrelation of hidden activations[28], soft orthogonality of weights[29] and the redundancy-reduction objectives of self-supervised learning[30,31]. Channel and spatial attention modules such as squeeze-and-excitation[32], CBAM[33] and attention gates[34] reweight features, but none of them asks whether a position is redundant with its neighbours.

Determinants ask exactly that question. In quantum mechanics, the Slater determinant of a set of one-particle orbitals vanishes when two fermions occupy the same state, which is the Pauli exclusion principle in algebraic form[35]. Neural networks that output Slater determinants now solve the electronic Schrödinger equation for small molecules[36,37]. In statistics and machine learning, the determinant of a Gram matrix is the kernel of determinantal point processes, which favour diverse subsets[38,39] and have been used to diversify training mini-batches[40]. Carrying this algebra into a segmentation network is tempting: a layer that silences positions whose features are linear combinations of their neighbours would, in principle, suppress repetitive mucosal texture.

Whether such a layer helps is an empirical question, and the usual way of answering it is weak. New modules are typically compared with a plain baseline, in one split and one training run. A gain observed this way may come from extra parameters, from generic gating, or from run-to-run noise, which in segmentation is often as large as the gains being reported[41]. Leakage between near-duplicate frames inflates scores further[42], and the choice and aggregation of metrics can change conclusions[43,44]. Varoquaux and Cheplygina documented how these habits slow progress in medical imaging[45].

We therefore designed the evaluation before the method could flatter itself. We built a Slater exclusion gate whose score is a Schur-complement distance with an exact zero set, and we compared it with two controls that share its parameters: a gate without any interaction between tokens, and a gate that excludes duplicates in pairs only. The hypotheses, their order and the decision rule were fixed in the analysis code before the first model was trained. Folds were grouped by perceptual hash, cross-validation was repeated with three seeds, and every model was tested without retraining on four external datasets (Fig. 1). The gate failed every prespecified hypothesis. We report this result in full, explain it with the gate's own diagnostics, and show that the same determinant, added to the loss as a penalty, reduces redundancy without loss of accuracy. We also release reimplementations of PraNet and

Polyp-PVT that reproduce the official code to the last bit, and a threshold-averaged Dice that matches the reference evaluation toolbox.

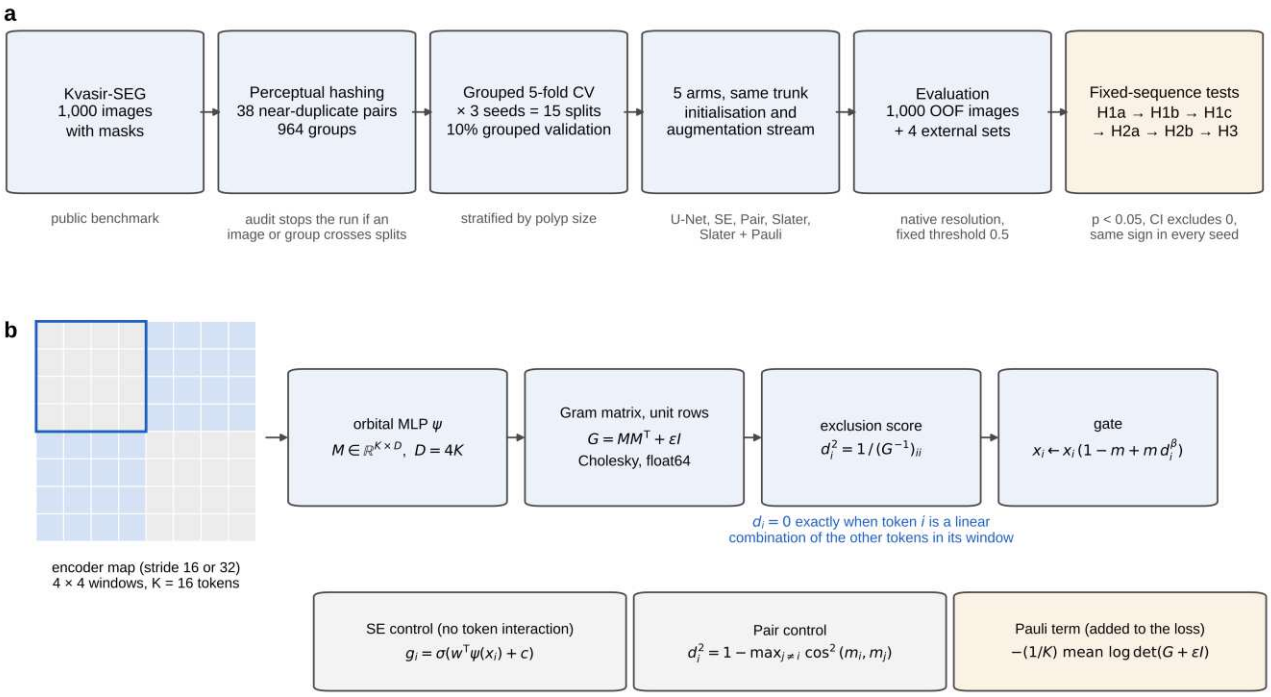


Figure 1. Study design and the Slater exclusion gate. (a) Evaluation pipeline. Near-duplicate frames were grouped before splitting, every arm was trained on the same 15 splits with identical initialisation of the shared network and identical augmented images, and the hypotheses were tested in a fixed order. (b) The gate. Tokens of each 4 × 4 window are mapped to orbital vectors; the Schur complement of the Gram matrix gives each token's distance to the span of the others, which is zero exactly when the token is linearly dependent on them. The two controls share the orbital network but remove the interaction between tokens (SE) or restrict it to pairs (Pair). The Pauli term is an optional loss on the same Gram matrix.

Results

Data audit and leakage control

All 1,000 Kvasir-SEG image–mask pairs were readable and size-matched. The bounding boxes released with the dataset agreed with the connected components of the masks (mean box IoU 0.983), and 48 images contained more than one polyp. Polyps covered a median of 11.4% of the image (5th to 95th percentile, 2.5% to 40.8%), and the automatic detector flagged 2.6% of pixels as specular highlights on average (Supplementary Table S1).

Perceptual hashing found 38 pairs of near-duplicate images, which formed 964 groups; 57 images belonged to a group of two or more, the largest holding 12 frames. In the 15 seed–fold splits, no image or group appeared on both sides of a split, and the smallest Hamming distance between a training image and a test image always exceeded the grouping threshold of

6 bits. No external image was a near-duplicate of a Kvasir-SEG image (minimum Hamming distance, 8 to 10 bits).

Verification of the implementation

The gate passed eleven executable checks of its stated properties, including antisymmetry under token exchange, the exclusion property, the Schur-complement identity, the Cauchy–Binet expansion, gradient correctness and exact masking of incomplete windows (Supplementary Table S2). Our PraNet and Polyp-PVT implementations gave outputs identical to those of the official code on GPU (maximum absolute difference exactly zero) for 352 × 352 and 256 × 320 inputs. Our threshold-averaged Dice and IoU gave the same values as a line-by-line translation of the PraNet MATLAB toolbox on all 100 test cases.

Main comparison

All five arms reached a per-image Dice between 0.900 and 0.905 (Table 2, Fig. 2a). The Slater gate (0.9006) was indistinguishable from the gate without the determinant (0.9007) and slightly below the plain U-Net (0.9038). The arm trained with the log-determinant penalty had the highest mean Dice (0.9054) and the smallest spread between folds (SD 0.0099). Fold-to-fold variation was larger than any difference between arms: within a single split, the Slater gate and its control differed by up to 0.036 Dice, in both directions (Fig. 2a).

Table 2. Kvasir-SEG, grouped five-fold cross-validation repeated with three seeds. Values are mean ± SD across seeds of the per-image metrics; HD95 is in native pixels.

Arm	Dice	IoU	Precision	Sensitivity	HD95	$Dice_{256}$	Dice 95% CI
U-Net	0.9038 ± 0.0001	0.8476 ± 0.0011	0.9030 ± 0.0015	0.9338 ± 0.0026	42.3 ± 0.5	0.8831	0.8955–0.9125
SE control	0.9007 ± 0.0048	0.8434 ± 0.0061	0.8982 ± 0.0073	0.9331 ± 0.0022	43.5 ± 1.0	0.8761	0.8923–0.9092
Pair control	0.8998 ± 0.0038	0.8411 ± 0.0048	0.8974 ± 0.0042	0.9315 ± 0.0030	43.4 ± 0.9	0.8744	0.8911–0.9082
Slater gate	0.9006 ± 0.0023	0.8429 ± 0.0037	0.8998 ± 0.0069	0.9330 ± 0.0058	44.0 ± 1.2	0.8771	0.8920–0.9093
Slater + Pauli	0.9051 ± 0.0010	0.8489 ± 0.0022	0.9060 ± 0.0072	0.9319 ± 0.0053	43.0 ± 0.2	0.8818	0.8972–0.9136

$Dice_{256}$, threshold-averaged Dice of the PraNet toolbox, mean of the 15 runs (14 for Slater + Pauli). Dice 95% CI, bootstrap over images (2,000 resamples) of the per-image Dice averaged across seeds. Specificity was 0.980–0.982 for all arms. These values come from five-fold cross-validation at 256 × 256 and are not directly comparable with Table 1.

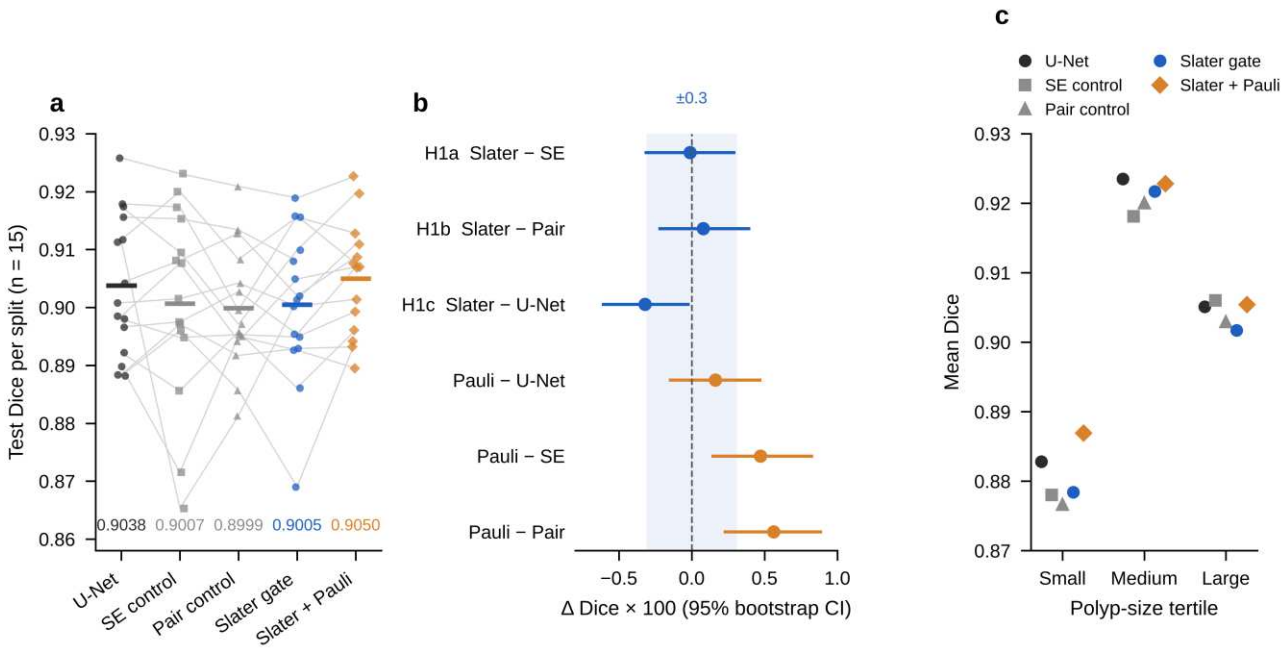


Figure 2. Main comparison on Kvasir-SEG. (a) Test Dice of each of the 15 seed–fold splits; grey lines join the same split across arms and horizontal bars give the mean. (b) Differences in per-image Dice with 95% bootstrap intervals (2,000 resamples, n = 1,000 images). The shaded band marks ±0.003; the interval of the primary comparison, H1a, lies inside it. Blue, prespecified hypotheses; orange, secondary comparisons of the Pauli arm. (c) Mean Dice by polyp-size tertile.

Accuracy depended on polyp size more than on the arm. All arms did worst on the smallest third of polyps (Dice 0.877 to 0.887) and best on medium-sized polyps (0.918 to 0.924), and within each tertile the spread between arms stayed below 0.011 (Fig. 2c).

Prespecified hypotheses

None of the six prespecified hypotheses was confirmed (Table 3). The primary hypothesis H1a failed: the Slater gate minus its matched control gave −0.0001 Dice (95% CI −0.0031 to 0.0029; Wilcoxon p = 0.999; Nadeau–Bengio p = 0.982). The interval is narrow enough to rule out any benefit larger than 0.003 Dice, so this is a precise null result rather than an underpowered one. Under the fixed-sequence procedure, the remaining hypotheses are exploratory. Two of them reached nominal significance on the per-image Wilcoxon test but failed the robustness criteria: the gain over the pairwise control (H1b, +0.0008) changed sign across seeds, and the loss against the plain U-Net (H1c, −0.0032) was not significant once training variance was accounted for (Nadeau–Bengio p = 0.417).

Table 3. Prespecified hypotheses, tested in fixed sequence. Differences are Slater gate minus comparator on per-image values averaged over seeds (n = 1,000).

Hypothesis	Versus	Metric	Difference (95% CI)	Wilcoxon p	Nadeau–Bengio p	Same sign	Outcome
H1a, effect of the determinant	SE	Dice	−0.0001 (−0.0031, 0.0029)	0.999	0.982	1/3	Not confirmed

Hypothesis	Versus	Metric	Difference (95% CI)	Wilcoxon p	Nadeau–Bengio p	Same sign	Outcome
H1b, many-token vs pairwise	Pair	Dice	+0.0008 (−0.0022, 0.0039)	<0.001	0.921	2/3	Exploratory
H1c, gain over baseline	U-Net	Dice	−0.0032 (−0.0061, −0.0003)	<0.001	0.417	0/3	Exploratory
H2a, highlight false positives	SE	FPR in highlights	+0.0017 (−0.0005, 0.0041)	0.991	0.778	1/3	Exploratory
H2b, false-alarm components	SE	Components per image	−0.0023 (−0.0123, 0.0073)	0.665	0.857	2/3	Exploratory
H3, token redundancy	SE	Mean |cos|	+0.0043 (0.0029, 0.0056)	<0.001	0.591	0/3	Exploratory

Confirmation required $p < 0.05$ in the expected direction (higher Dice; lower false-positive rate, components and redundancy), a 95% bootstrap interval excluding zero and the same sign in each of the three seeds. Because H1a failed, H1b to H3 are exploratory. 'Same sign' counts the seeds in which the difference went in the expected direction.

Why the gate fails

The gate values explain the null result (Fig. 3a). At the end of training, the mean Slater gate was 0.399 at stride 16 and 0.078 at stride 32, against 0.956 and 0.462 for the control without the determinant. The gate therefore judged nearly every deep token redundant. This is what the algebra predicts for smooth feature maps: inside a 4 × 4 window at stride 32, neighbouring tokens are close to linearly dependent, so the Schur-complement distance is small for all of them at once. A gate that scales the whole map by a similar factor carries no information about where an artefact is, and batch normalisation in the following layers can largely absorb it.

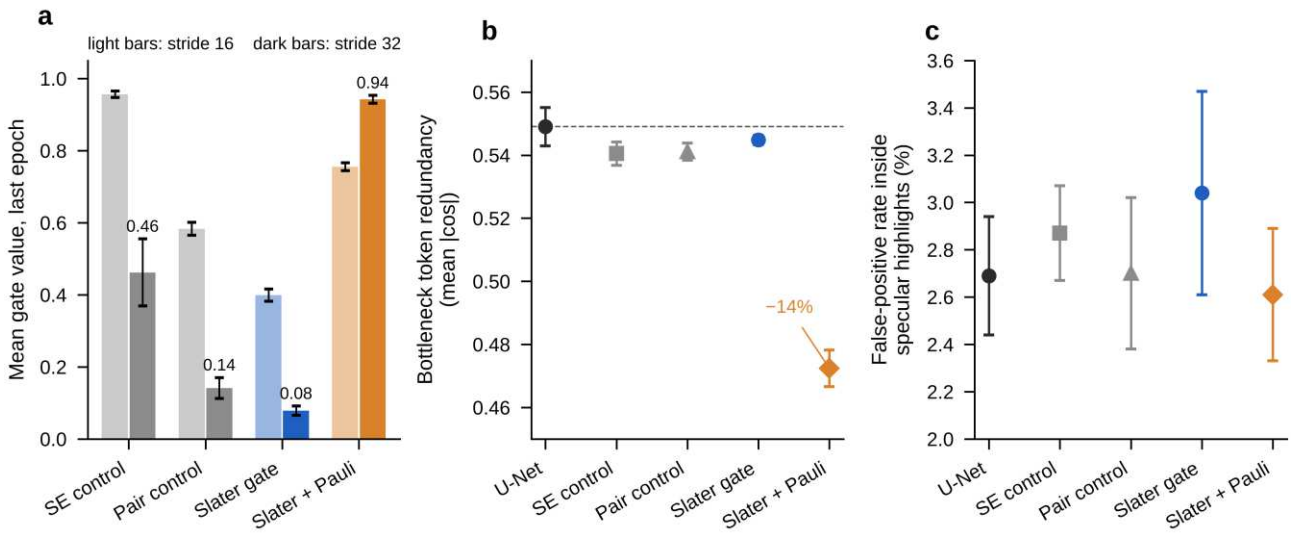


Figure 3. Mechanism. (a) Mean gate value at the last epoch for the two gated stages (light bars, stride 16; dark bars, stride 32); error bars show SD over runs. The Slater gate falls to 0.08 at stride 32. (b) Redundancy of bottleneck tokens, measured as the mean absolute cosine similarity between distinct tokens (mean ± SD over seeds); the dashed line marks the U-Net. (c) False-positive rate inside automatically detected specular highlights (default detector setting; mean ± SD over seeds).

The gate also failed the mechanistic hypotheses. Inside specular highlights, its false-positive rate was 3.04% against 2.87% for the control (H2a; Fig. 3c), and the direction was the same

for all three detector settings (+0.0016, +0.0017 and +0.0017; every interval included zero). The number of false-alarm components per image, which counts bubbles and mucus as well as highlights, did not change (H2b; −0.0023, 95% CI −0.0123 to 0.0073). Token redundancy moved the wrong way, by +0.8% (H3).

A log-determinant penalty reduces redundancy

Adding the Pauli term to the loss of the gated model changed the behaviour of the same quantity. Mean similarity between bottleneck tokens fell from 0.549 for the U-Net and 0.545 for the Slater gate to 0.472, a reduction of 14% relative to the U-Net with a between-seed SD of 0.006 (Fig. 3b). The penalty also reopened the gate (0.942 at stride 32). Dice stayed at the level of the U-Net (+0.0016, 95% CI −0.0015 to 0.0047) and exceeded the pairwise control (+0.0056, 95% CI 0.0023 to 0.0088; Holm-adjusted $p < 0.001$). These comparisons were secondary, so we read them as exploratory. Lower redundancy did not translate into fewer false positives: the Pauli arm's false-positive rate inside highlights (2.61%) was within one SD of the U-Net's (2.69%).

Generalisation to other centres

Without retraining, every arm lost accuracy on external data (Fig. 4). The drop was largest on ETIS-LaribPolypDB (Dice 0.653 to 0.684) and CVC-ColonDB (0.720 to 0.735). The Slater gate never differed from its matched control by more than 0.003 Dice. Only on CVC-300 did that difference survive Holm correction (+0.0009, adjusted $p = 0.023$), a gap far smaller than the spread between datasets. On ETIS-LaribPolypDB the gate fell below the U-Net by 0.018 (95% CI −0.026 to −0.010). No arm exceeded the plain U-Net by more than 0.001 Dice on any external set.

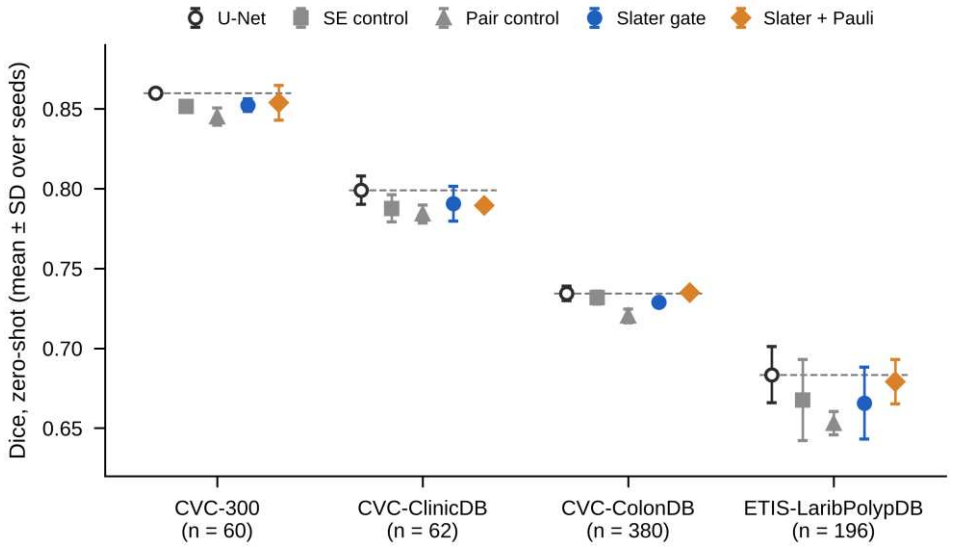


Figure 4. Zero-shot generalisation. Models trained on Kvasir-SEG were applied without retraining to four external test sets. Each mark is the mean over three seeds of the per-image Dice averaged across the five fold models; error bars show SD across seeds. Dashed lines mark the U-Net.

Discussion

A gate built on the Slater determinant, inserted at the two deepest stages of a U-Net, did not improve polyp segmentation. The comparison that isolates the determinant, against a gate with the same orbital network and no interaction between tokens, gave a difference of −0.0001 Dice with an interval of ±0.003. The gate also left false positives inside highlights unchanged and did not lower redundancy. We regard the interval width as the main result: it bounds the benefit of the determinant in this setting more tightly than the fold-to-fold noise that many single-run comparisons rely on.

The gate's own values account for the failure. The exclusion score is exact, and its zero set is exactly the set of tokens that are linear combinations of their neighbours. Deep feature maps vary slowly in space, so most tokens in a 4 × 4 window come close to that set. The gate then attenuates almost everything (mean value 0.08 at stride 32), which amounts to rescaling the whole map. Linear dependence is a property of the representation; an artefact is a property of the image; our data show that, in these features, the first does not track the second. A gate would need a reference against which redundancy is judged, such as a learned prior or a calibration of the score within each window, before it could single out artefacts.

Added to the loss as a penalty, the determinant behaved quite differently. The Pauli term cut inter-token similarity by 14% and kept Dice at the level of the best arm, in line with earlier decorrelation and orthogonality penalties[28,29] and with the diversity objective of determinantal point processes[39]. It did not reduce false positives or improve external generalisation. Lower

redundancy is therefore not, on its own, a route to more robust segmentation, a point that is easy to assume and seldom tested.

The experiment also says something about how such modules should be evaluated. Fold-level differences between the gate and its control reached 0.036 Dice in either direction. A single run against a plain baseline could have reported a gain or a loss of that size. The matched controls, three repetitions of grouped cross-validation and the Nadeau–Bengio correction were what made the null result visible. These controls follow existing recommendations for medical imaging and machine-learning benchmarks[41,42,45], and we suggest them as a minimum for any claim about mechanism.

Our study has limitations. All models were trained on one dataset. The external sets were the public test subsets distributed with PraNet, two of which hold about 60 images, so their intervals are wide. The specular detector is a colour-threshold heuristic that we did not validate against annotated highlights, although the direction of H2a was the same for three thresholds and H2b does not depend on the detector. We tested 4 × 4 windows only; shifted windows, a global gate and the square case D = K are implemented in the released code but were not part of the prespecified comparison. Patient identifiers are not public, so perceptual hashing groups only visually similar frames. Finally, the gate was inserted in a ResNet-34 U-Net; whether stronger encoders such as PVTv2 respond differently is open, and the released PraNet and Polyp-PVT implementations allow this to be tested under the same protocol.

In conclusion, an exact algebraic test of redundancy did not make a segmentation network more accurate or more robust to endoscopic artefacts. The gate's guarantee holds, but deep features meet its condition almost everywhere. Used as a soft penalty, the same determinant reduces redundancy by 14% at no cost in accuracy. We hope that the released code, splits and controls make it easier to test the next quantum-inspired module against the comparison that matters.

Methods

Datasets

Kvasir-SEG contains 1,000 colonoscopy images from Vestre Viken Health Trust, Norway, with widths from 332 to 1,920 pixels and heights from 352 to 1,072 pixels, binary polyp masks and bounding boxes[4]. The masks are stored as JPEG files; we binarised them at an intensity of 127 after confirming that intermediate values occurred only at edges. For external testing we used the test sets distributed with PraNet[9]: CVC-ClinicDB (62 images)[5], CVC-

ColonDB (380)[6], ETIS-LaribPolypDB (196)[7] and CVC-300 from EndoScene (60)[8]. All datasets are public and de-identified, so no ethics approval was required.

Preprocessing and augmentation

Images were resized to 256 × 256 pixels with area interpolation. Specular highlights were detected as pixels whose HSV saturation was below 0.25 and whose value exceeded the larger of 0.85 and the image's 97th percentile, dilated by one pixel and removed with Telea inpainting[46]. Contrast was then enhanced with CLAHE[47] on the L* channel of CIELAB (clip limit 2.0, 8 × 8 tiles), and inputs were normalised with ImageNet statistics. Training augmentation used Albumentations[48]: flips, 90° rotations, affine transforms (±30°, scale 0.85 to 1.15, translation ±6%), elastic or grid distortion, brightness, contrast, gamma and hue shifts, and Gaussian or motion blur. The augmentation seed of each sample was a function of the run seed, the epoch and the image index, so all arms trained in the same seed and fold saw identical augmented images in the same order.

Slater exclusion gate

A feature map from the encoder is divided into non-overlapping windows of w × w positions (w = 4), each holding $K = w^2 = 16$ tokens $x_1, \ldots, x_K$. A shared two-layer perceptron ψ maps each token to D orbital values, giving the K × D matrix M with entries $M_{i\alpha} = \psi_\alpha(x_i)$. After normalising each row of M to unit length (written $\hat{M}$), the Gram matrix of the window and its determinant are

$$G = \hat{M}\hat{M}^{\mathsf{T}}, \qquad \det G = \sum_{S \subset \{1,\ldots,D\},\ |S| = K} \det\left(\hat{M}_{:,S}\right)^2 \tag{1}$$

By the Cauchy–Binet formula, det G is the sum of the squared Slater determinants obtained by choosing K occupied orbitals among D; for D = K it reduces to a single squared Slater determinant[35]. Exchanging two tokens flips the sign of every term but leaves det G unchanged. To score individual tokens, we use the Schur complement of token i in G:

$$d_i^2 = \frac{\det G}{\det G_{\overline{ii}}} = \frac{1}{(G^{-1})_{ii}} = \operatorname{dist}^2\left(\hat{m}_i,\ \operatorname{span}\{\hat{m}_j\}_{j \neq i}\right) \tag{2}$$

Proposition 1 follows directly: $d_i = 0$ if and only if the orbital vector of token i is a linear combination of the other vectors in its window. This is the only property we claim for the gate. Whether artefacts produce such dependence is what hypotheses H2 and H3 test. The gate rescales each token as

$$x_i' = x_i\left(1 - m + m\, d_i^{\beta}\right), \qquad m = \sigma(\mu), \quad \beta = \mathrm{softplus}(b) \tag{3}$$

with μ and b learned and initialised so that m = 0.27 and β = 1. All gate quantities are computed in double precision from a Cholesky factorisation of G + εI with ε = 10^{-4}. When the map size is not a multiple of w, padded positions receive vectors orthogonal to all others; G becomes block-diagonal and the scores of real tokens equal those of the real subset exactly. We set D = 4K = 64. For D = K, the expected distance of a random unit vector to the span of K − 1 others is close to $1/\sqrt{K}$, so the gate would attenuate every token; for D = 4K the measured mean distance was 0.87.

Controls, regulariser and network

The two controls share the orbital perceptron and the update rule of Eq. (3). The SE control reads each token's orbital vector through a linear layer and a sigmoid, so tokens do not interact[32]. The Pair control uses $d_i^2 = 1 - \max_{j \neq i} \cos^2(\hat{m}_i, \hat{m}_j)$, the two-token case of Eq. (2), which detects duplicates but not combinations. The Pauli regulariser adds $\lambda\, L_{\mathrm{Pauli}}$ to the loss, with λ = 0.01 (Eq. 4). All arms use a U-Net[10] with an ImageNet-pretrained ResNet-34 encoder[49] and gates after the encoder stages of stride 16 (256 channels) and stride 32 (512 channels). Gates were created after the encoder and decoder, so that within a seed and fold every arm started from identical trunk weights. The U-Net has 24.4 M parameters and each gated model 24.5 M; the three gated models differ by fewer than 200 parameters.

Training

The segmentation loss combined binary cross-entropy (weight 0.5) with the Focal Tversky loss[50] (α = 0.7, β = 0.3, γ = 0.75):

$$\mathcal{L} = 0.5\,\mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{FT}} + \lambda\,\mathcal{L}_{\mathrm{Pauli}}, \qquad \mathcal{L}_{\mathrm{Pauli}} = -\frac{1}{K}\overline{\log\det\left(G + \epsilon I\right)} \tag{4}$$

We trained with AdamW[51] (learning rate 3 × 10^{-4}, weight decay 10^{-4}, batch size 16) for at most 60 epochs, with a three-epoch linear warm-up, cosine decay, gradient-norm clipping at 1.0 and mixed precision (the gate itself in double precision). Early stopping monitored Dice on an internal validation set made of 10% of the training groups, with a patience of 15 epochs. At inference we averaged predictions over the identity and horizontal and vertical flips. The decision threshold was fixed at 0.5 and never tuned.

Validation design

We computed a 64-bit DCT perceptual hash for every image and linked pairs within a Hamming distance of 6 into groups. StratifiedGroupKFold then produced five folds stratified by polyp-size tertile, and the procedure was repeated with seeds 42, 1337 and 2024. An executable audit stopped the pipeline if an image or a group was shared between training, validation and test, or if the smallest training–test Hamming distance fell to the grouping threshold. Each image was therefore tested exactly once per seed.

Evaluation metrics

Predicted logits were resampled to the native resolution of each image and thresholded at 0.5. Following Metrics Reloaded[43,44], we computed per-image Dice, IoU, sensitivity, specificity, precision and the 95th-percentile Hausdorff distance. The false-positive rate inside highlights is the fraction of non-polyp highlight pixels predicted as polyp, computed for three detector settings (saturation below 0.20, 0.25 or 0.30 with value above 0.90, 0.85 or 0.80). A false-alarm component is a predicted connected component that does not touch the ground truth and covers at least 0.05% of the image. Redundancy is the mean absolute cosine similarity between distinct tokens of the bottleneck. For comparison with published work, we also computed the threshold-averaged Dice and IoU of the PraNet toolbox: each probability map is normalised to [0, 1], quantised to 8 bits and binarised at 256 thresholds from 1 to 0, and an image scores zero at a threshold that yields no true positive[9].

Statistical analysis

The hypotheses, their order and the decision rule were written into the analysis code before any model was trained. The primary endpoint was the per-image Dice averaged over the three seeds (n = 1,000). We tested H1a (Slater gate vs SE control), H1b (vs Pair control), H1c (vs U-Net), H2a (false-positive rate inside highlights), H2b (false-alarm components per image) and H3 (token redundancy) in that fixed sequence at $\alpha = 0.05$, stopping at the first failure. A hypothesis was confirmed only if a two-sided Wilcoxon signed-rank test[52] gave $p < 0.05$ in the expected direction, the 95% bootstrap interval of the mean difference[53] (2,000 resamples) excluded zero, and the difference had the same sign in every seed. As sensitivity analyses, we applied the corrected resampled t-test of Nadeau and Bengio[54] to the 15 seed–fold units and a linear mixed model with a random intercept for each image. All other comparisons were corrected with the Holm procedure[55], following common practice for comparisons across many models[56]. Effects are reported as mean differences with bootstrap intervals and rank-biserial correlations.

Reimplementation of PraNet and Polyp-PVT

Because the official PraNet and Polyp-PVT repositories carry no software licence, we wrote independent implementations with the same module names. This allowed us to load the official weights and compare outputs. With the official weights transferred, all four PraNet outputs and both Polyp-PVT outputs matched the official code exactly on random inputs of 352 × 352 and 256 × 320 pixels, and the PVTv2-B2[23] and Res2Net-50[15] backbones obtained through timm gave the same features as the official ones. These implementations are not used in the comparisons above; we release them so that the gate can be tested on stronger architectures under the same protocol.

Computation

Experiments ran on a single NVIDIA T4 GPU through Kaggle notebooks. The 74 training runs took 10.4 h in total (8.4 min per run on average). At 256 × 256 pixels and batch size 1, inference took between 6.5 and 9.6 ms per image for all five arms, and the differences between arms were within measurement noise.

Data availability

Kvasir-SEG is available from the Simula Research Laboratory (https://datasets.simula.no/kvasir-seg/). The PraNet test sets are available from the PraNet repository (https://github.com/DengPingFan/PraNet). Fold assignments for all seeds (Supplementary Data 1), per-image metrics for every run (Supplementary Data 2) and training histories (Supplementary Data 3) are provided with this article; their contents are described in Supplementary Table S4.

Acknowledgements

We thank the creators of Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB and EndoScene for making their data public, and the authors of PraNet and Polyp-PVT for releasing their code and test sets.

Author contributions

A.B. conceived and designed the study, wrote the software, ran the experiments, analysed the data, prepared the figures and wrote the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The author declares no competing interests.

References

1. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Corley, D. A. et al. Adenoma detection rate and risk of colorectal cancer and death. *N. Engl. J. Med.* **370**, 1298–1306 (2014).
3. Zhao, S. et al. Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: a systematic review and meta-analysis. *Gastroenterology* **156**, 1661–1674 (2019).
4. Jha, D. et al. Kvasir-SEG: a segmented polyp dataset. In *MultiMedia Modeling (MMM 2020)*, Lecture Notes in Computer Science Vol. 11962, 451–462 (Springer, 2020).
5. Bernal, J. et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **43**, 99–111 (2015).
6. Tajbakhsh, N., Gurudu, S. R. & Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* **35**, 630–644 (2016).
7. Silva, J., Histace, A., Romain, O., Dray, X. & Granado, B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **9**, 283–293 (2014).
8. Vázquez, D. et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* **2017**, 4037190 (2017).
9. Fan, D.-P. et al. PraNet: parallel reverse attention network for polyp segmentation. In *MICCAI 2020*, Lecture Notes in Computer Science Vol. 12266, 263–273 (Springer, 2020).
10. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, Lecture Notes in Computer Science Vol. 9351, 234–241 (Springer, 2015).
11. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **39**, 1856–1867 (2020).
12. Jha, D. et al. ResUNet++: an advanced architecture for medical image segmentation. In *Proc. IEEE International Symposium on Multimedia (ISM)*, 225–230 (IEEE, 2019).
13. Jha, D. et al. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. *IEEE J. Biomed. Health Inform.* **25**, 2029–2040 (2021).
14. Fang, Y., Chen, C., Yuan, Y. & Tong, K. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *MICCAI 2019*, Lecture Notes in Computer Science Vol. 11764, 302–310 (Springer, 2019).
15. Gao, S.-H. et al. Res2Net: a new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 652–662 (2021).
16. Zhao, X., Zhang, L. & Lu, H. Automatic polyp segmentation via multi-scale subtraction network. In *MICCAI 2021*, Lecture Notes in Computer Science Vol. 12901, 120–130 (Springer, 2021).
17. Wei, J. et al. Shallow attention network for polyp segmentation. In *MICCAI 2021*, Lecture Notes in Computer Science Vol. 12901, 699–708 (Springer, 2021).

18. Kim, T., Lee, H. & Kim, D. UACANet: uncertainty augmented context attention for polyp segmentation. In *Proc. 29th ACM International Conference on Multimedia*, 2167–2175 (ACM, 2021).
19. Lou, A., Guan, S., Ko, H. & Loew, M. H. CaraNet: context axial reverse attention network for segmentation of small medical objects. In *Proc. SPIE Medical Imaging 2022: Image Processing* **12032**, 120320D (2022).
20. Huang, C.-H., Wu, H.-Y. & Lin, Y.-L. HarDNet-MSEG: a simple encoder–decoder polyp segmentation neural network that achieves over 0.9 mean Dice and 86 FPS. Preprint at https://arxiv.org/abs/2101.07172 (2021).
21. Zhang, Y., Liu, H. & Hu, Q. TransFuse: fusing transformers and CNNs for medical image segmentation. In *MICCAI 2021*, Lecture Notes in Computer Science Vol. 12901, 14–24 (Springer, 2021).
22. Dong, B. et al. Polyp-PVT: polyp segmentation with pyramid vision transformers. *CAAI Artif. Intell. Res.* **2**, 9150015 (2023).
23. Wang, W. et al. PVT v2: improved baselines with pyramid vision transformer. *Comput. Vis. Media* **8**, 415–424 (2022).
24. Ji, G.-P. et al. Video polyp segmentation: a deep learning perspective. *Mach. Intell. Res.* **19**, 531–549 (2022).
25. Mei, J. et al. A survey on deep learning for polyp segmentation: techniques, challenges and future trends. Preprint at https://arxiv.org/abs/2311.18373 (2023).
26. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
27. Arnold, M., Ghosh, A., Ameling, S. & Lacey, G. Automatic segmentation and inpainting of specular highlights for endoscopic imaging. *EURASIP J. Image Video Process.* **2010**, 814319 (2010).
28. Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L. & Batra, D. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations (ICLR)* (2016).
29. Bansal, N., Chen, X. & Wang, Z. Can we gain more from orthogonality regularizations in training deep networks? In *Advances in Neural Information Processing Systems* **31** (2018).
30. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: self-supervised learning via redundancy reduction. In *Proc. 38th International Conference on Machine Learning*, PMLR **139**, 12310–12320 (2021).
31. Bardes, A., Ponce, J. & LeCun, Y. VICReg: variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)* (2022).
32. Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 2011–2023 (2020).
33. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: convolutional block attention module. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science Vol. 11211, 3–19 (Springer, 2018).
34. Oktay, O. et al. Attention U-Net: learning where to look for the pancreas. Preprint at https://arxiv.org/abs/1804.03999 (2018).
35. Slater, J. C. The theory of complex spectra. *Phys. Rev.* **34**, 1293–1322 (1929).
36. Pfau, D., Spencer, J. S., Matthews, A. G. D. G. & Foulkes, W. M. C. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2**, 033429 (2020).
37. Hermann, J., Schätzle, Z. & Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.* **12**, 891–897 (2020).
38. Macchi, O. The coincidence approach to stochastic point processes. *Adv. Appl. Probab.* **7**, 83–122 (1975).
39. Kulesza, A. & Taskar, B. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.* **5**, 123–286 (2012).
40. Zhang, C., Kjellström, H. & Mandt, S. Determinantal point processes for mini-batch diversification. In *Proc. 33rd Conference on Uncertainty in Artificial Intelligence (UAI)* (2017).
41. Bouthillier, X. et al. Accounting for variance in machine learning benchmarks. In *Proc. Machine Learning and Systems* **3** (2021).

42. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).
43. Maier-Hein, L. et al. Metrics reloaded: recommendations for image analysis validation. *Nat. Methods* **21**, 195–212 (2024).
44. Reinke, A. et al. Understanding metric-related pitfalls in image analysis validation. *Nat. Methods* **21**, 182–194 (2024).
45. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit. Med.* **5**, 48 (2022).
46. Telea, A. An image inpainting technique based on the fast marching method. *J. Graph. Tools* **9**, 23–34 (2004).
47. Zuiderveld, K. Contrast limited adaptive histogram equalization. In *Graphics Gems IV* (ed. Heckbert, P. S.) 474–485 (Academic Press, 1994).
48. Buslaev, A. et al. Albumentations: fast and flexible image augmentations. *Information* **11**, 125 (2020).
49. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
50. Abraham, N. & Khan, N. M. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In *Proc. IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, 683–687 (2019).
51. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)* (2019).
52. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80–83 (1945).
53. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979).
54. Nadeau, C. & Bengio, Y. Inference for the generalization error. *Mach. Learn.* **52**, 239–281 (2003).
55. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
56. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BougarneSupplementaryInformation.docx