THE LANCET
Digital Health

Supplementary appendix 3

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Wagner SK, Liefers B, Radia M, et al. Development and international validation of custom-engineered and code-free deep-learning models for detection of plus disease in retinopathy of prematurity: a retrospective study. *Lancet Digit Health* 2023; published online April 21. https://doi.org/10.1016/S2589-7500(23)00050-X.

Appendix 3

Automated detection of plus disease in retinopathy of prematurity using deep learning: A retrospective cohort study

Siegfried K. Wagner MD[1,2,3] *, Bart Liefers PhD[1]*, Meera Radia MD[3], Gongyu Zhang MSc[1], Robbert Struyven MD[1,2,3], Livia Faes MD[1,3], Jonathan Than MD[3], Shafi Balal MD[3], Charlie Hennings MD[3], Caroline Kilduff MD[3], Pakinee Pooprasert MD[3], Sophie Glinton PhD[1], Meena Arunakirinathan MD[3], Periklis Giannakis MD[4], Imoro Zeba Braimah MD[5], Islam SH Ahmed PhD[6,7], Mariam Al-Feky PhD[8,9], Hagar Khalid PhD[3,10], Daniel Ferraz PhD[2,11], Juliana Vieira MD[12], Rodrigo Jorge PhD[12 #], Shahid Husain MD[13,14], Janette Ravelo BSc[13], Anne-Marie Hinds MD[3], Robert Henderson MD[15,16], Himanshu I. Patel MD[3,17], Susan Ostmo MS[18], J Peter Campbell MD[18] , Nikolas Pontikos PhD[1,2,3], Praveen J. Patel MD[1,2,3], Pearse A. Keane MD [1,2,3 #], Gill Adams MD[1,3], Konstantinos Balaskas MD[1,2,3]

[1] NIHR Moorfields Biomedical Research Centre, London, UK
[2] Institute of Ophthalmology, University College London, London, UK,
[3] Moorfields Eye Hospital NHS Foundation Trust, London, UK
[4] Institute of Health Sciences Education, Queen Mary University of London, London, UK
[5] Lions International Eye Centre, Korle-Bu Teaching Hospital, Accra, Ghana
[6] Faculty of Medicine, Alexandria University, Alexandria, Egypt
[7] Alexandria University Hospital, Alexandria, Egypt
[8] Department of Ophthalmology, Ain Shams University Hospitals, Cairo, Egypt
[9] Watany Eye Hospital, Cairo, Egypt
[10] Tanta University, Tanta, Egypt
[11] D'Or Institute for Research and Education, Sao Paulo, Brazil
[12] Department of Ophthalmology, RIbeirão Preto Medical School, University of São Paulo, RIbeirão Preto, Brazil
[13] Neonatology Department, Homerton University Hospital NHS Foundation Trust, London, UK.
[14] The Blizard Institute, Queen Mary University of London, London, UK
[15] Clinical and Academic Department of Ophthalmology, Great Ormond Street Hospital for Children, London, UK
[16] UCL-GOSH Institute of Child Health, University College London, London, UK
[17] The Royal London Hospital, Barts Health NHS Trust, London, UK
[18] Department of Ophthalmology, Oregon Health & Science University, Portland, OR, USA

Methods

Dataset and participants

Homerton University Hospital Trust provides tertiary-level neonatal care to the North Central and East London Neonatal Network of the London Neonatal Operational Delivery Network and serves an ethnically and socioeconomically diverse population within East London, UK. Our cohort reflected the rich ethnic diversity of the catchment population(1). Around 120-150 infants are admitted for ROP screening and treatment annually (2). Individual-level data for socioeconomic status was not available, however, the Homerton catchment population experiences among the highest level of socioeconomic deprivation when using the Index of Multiple Deprivation 2019 (1,3), a widely used measure of relative deprivation across seven domains (income, employment, education, health, and barriers to housing and services, crime and living environment) within the UK based on postcode. To maintain robust privacy preservation, images were exported in a fully anonymised form without any associated clinical metadata with the permission of the Caldicott Guardian at Homerton University Hospital.

Image grading

For grading, images were stored on encrypted folders housed within the Moorfields Eye Hospital Reading Centre and viewed using the open-source software ImageJ (4). Regarding where there were multiple images for the same patient, the ophthalmologist was not aware which image pertained to which patient. Junior ophthalmologists were required to complete the RCOphth e-learning for health module Eye-sight module on ROP (5) and American Academy of Ophthalmology Retinopathy of Prematurity Case-based training (6) prior to participating in grading.

Training

Bespoke

For the development of the bespoke model, the architecture was augmented to handle 512 x 512 input, by including an additional MaxPooling layer and a 1x1 convolution layer with 512 filters before the output layer. The model was trained using the Adam optimiser with cross-entropy as the loss function, a learning rate of 10^{-5} and batch size of 6 (randomly sampling 2 normal, 2 pre-plus and 2 plus images from the training set). During training, images were augmented by applying small random transformations (translation, rotation, scaling, horizontal flipping, brightness, contrast and gamma-corrections). Data augmentation was applied to avoid overfitting. Data augmentation consisted of a random transformation applied to the image. Spatially, this consisted of a rotation of -30 to 30 degrees, horizontal and vertical translation of up to 64 pixels and scaling up or down with a factor of up to 1.1. Pixel intensities (luminance values) were transformed using the formula c * I ^ g + b, with c the contrast (1/1.1 to 1.1), b the brightness (-0.1 to 0.1) and g gamma transform (1/1.5 to 1.5).

Training was continued until the model converged, which was monitored every 100 iterations on the tuning set. Convergence occurred after 7100-9500 iterations depending on the fold.

Code-free deep learning

For the CFDL model, images were uploaded to a secure storage ‘bucket’ within the Moorfields Research Informatics Cloud environment in conjunction with a comma-separated-value file indicating the file path, dataset allocation (e.g. train, tune, test) and corresponding class.

Evaluation

We developed saliency maps based on five separate techniques - XRAI heatmaps, Vanilla gradient, GradCAM, SmoothGRAD and Integrated Gradients though we note that such illustrations should be interpreted with caution(7–12).

Statistical analysis

For interrater reliability between more than two graders, we used two-way (all images were graded by the same set of clinicians) random effects consistency (priority for demonstrating that grades are similar in rank order among clinicians) average-measures intraclass correlation coefficient (ICC) (13–16)

Results

Misclassification audit

For the bespoke model, there were 24 disagreements between the model output and the reference standard within the internal validation dataset, of which 13 were pre-plus, seven as normal and four as plus. All plus misclassifications were labeled as pre-plus. Of the pre-plus cases, seven were misclassified as plus however six were considered normal by the bespoke model. In four of these six, one of the senior paediatric ophthalmologists had also considered the image normal. ,Of the 35 misclassifications by the CFDL model, 33 had a reference standard of pre-plus and two plus. The two plus cases were labeled as pre-plus however 22 of the pre-plus cases were misclassified as normal. Visual inspection highlighted some pre-plus images misclassified as plus by one of the models to have borderline features (e.g. severe pre-plus).

Rater	Role	Ophthalmology experience (years)
CR1	Consultant	41
CR2	Consultant	25
CR3	Consultant	21
CR4	Consultant	12
JR1	Paediatric Ophthalmology Fellow	7
AHP1	Specialist ROP Nurse	8
JR2	Resident	3
JR3	Resident	3
JR4	Resident	3
JR5	Resident	3

Supplementary Table 1: Role and level of experience of graders participating in the internal test set. Note that the reference standard was the majority vote of CR1, CR2 and CR3.

ROP: Retinopathy of prematurity.

		i-ROP	Brazil	Egypt Retcam	Egypt 3nethra
Images		100	92	45	101
Patients		70	46	45	33
Female sex		42	24	25	*
Ethnic Group	**Black**	13	1	-	-
	Middle Eastern	-	-	45	33
	White	80	37	-	-
	Other	7	8	-	-
Class[1]	**Normal**	54	72	13	30
	Pre-plus	31	20	32	71
	Plus	15			
Imaging device		Natus Medical Retcam	Natus Medical Retcam	Natus Medical Retcam	Forus Health 3nethra
Reference standard		Combiend classification of three expert ROP image graders and actual clinical diagnosis	BIO and image grading by single paediatric ophthalmologist	BIO and image grading by single paediatric ophthalmologist	BIO and image grading by single paediatric ophthalmologist

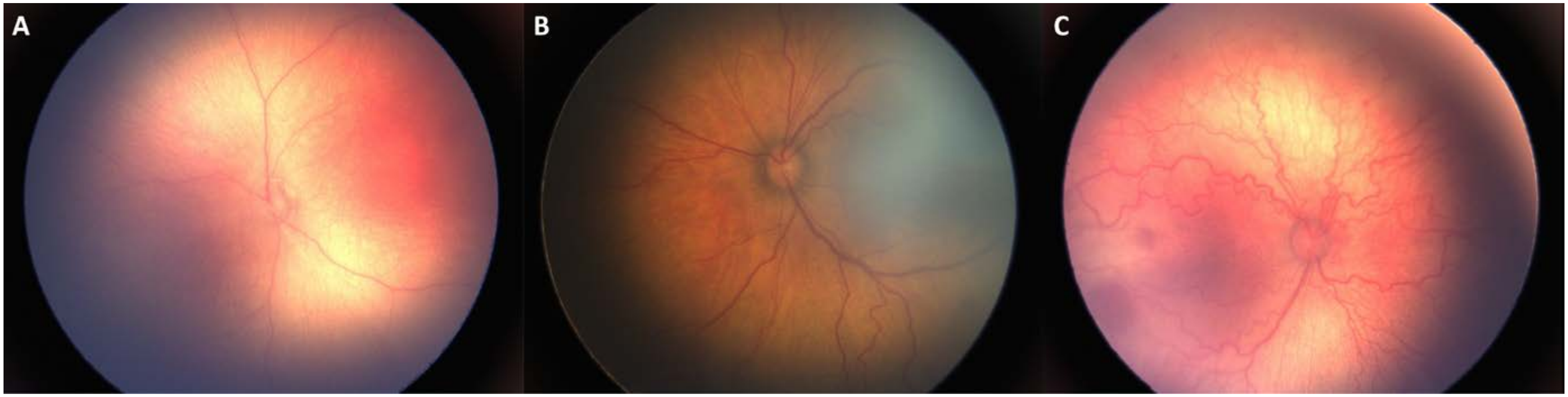
Supplementary Table 2: Characteristics of the external validation datasets. [1]Note that for disease class, images from Brazil and Egypt were graded in a binary fashion as presence of pre-plus/plus or normal. *Not reported. Data fully anonymised without biological sex data.

Dataset	Class		Normal	Pre-plus	Plus
Internal test set	Bespoke	Model 1	0·979	0·894	0·958
		Model 2	0·973	0·902	0·961
		Model 3	0·981	0·913	0·964
		Model 4	0·979	0·918	0·969
		Model 5	0·982	0·912	0·973
		Average of models	0·979 +/- 0·003	0·908 +/- 0·008	0·965 +/- 0·005
		Ensemble	0·986	0·927	0·974
	CFDL	Model	0·989	0·932	0·988
i-ROP external test set	Bespoke	Model 1	1	0·905	0·952
		Model 2	0·980	0·926	0·969
		Model 3	0·994	0·878	0·965
		Model 4	0·997	0·941	0·962
		Model 5	0·990	0·828	0·967
		Average of models	0·992 +/- 0·007	0·896 +/- 0·040	0·963 +/- 0·060
		Ensemble	1	0·942	0·976
	CFDL	Model	0·995	0·808	0·989

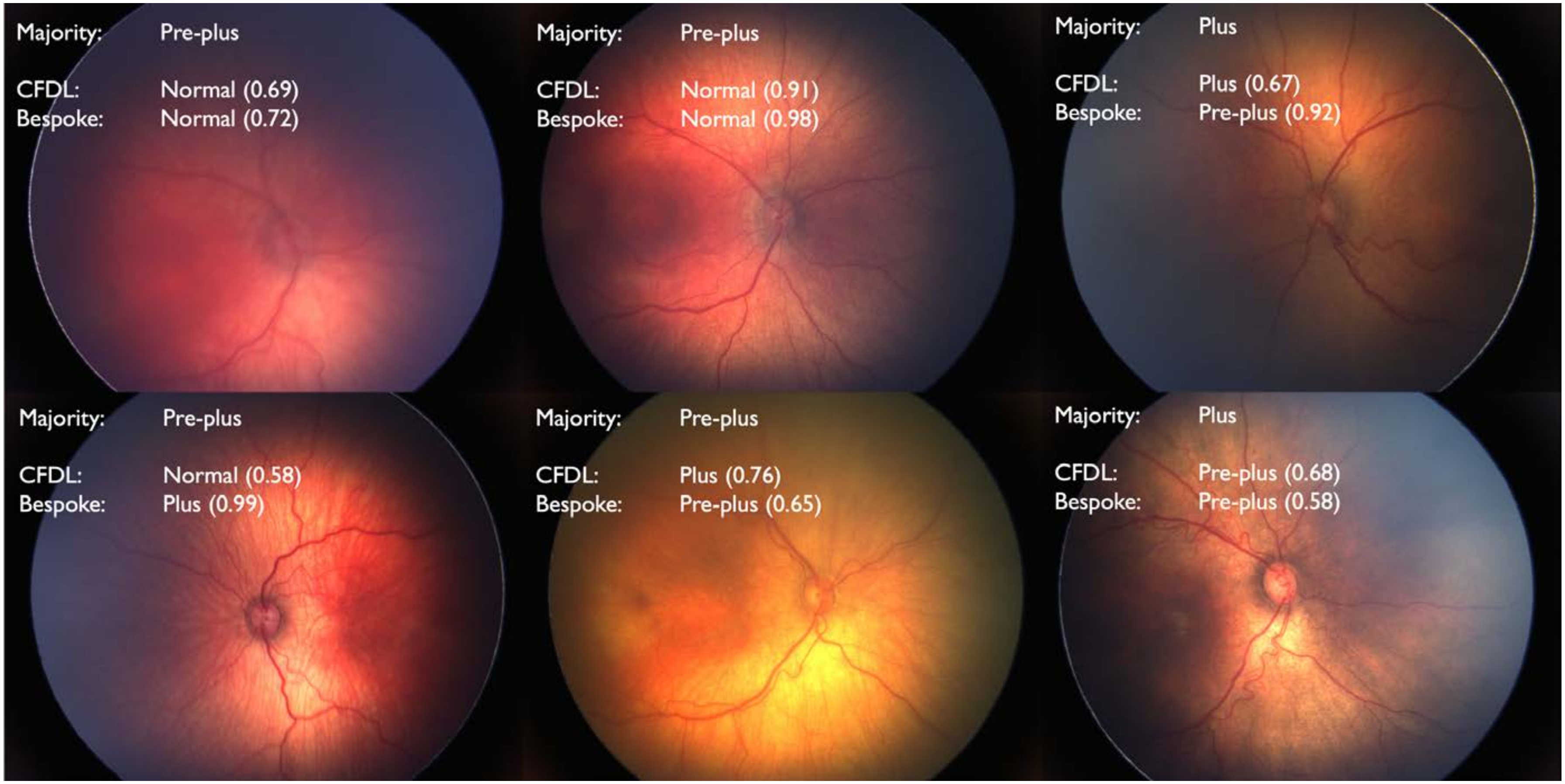
Supplementary Table 3: Performance of individual models for the bespoke model versus the ensemble and the CFDL model.

CFDL: code-free deep learning

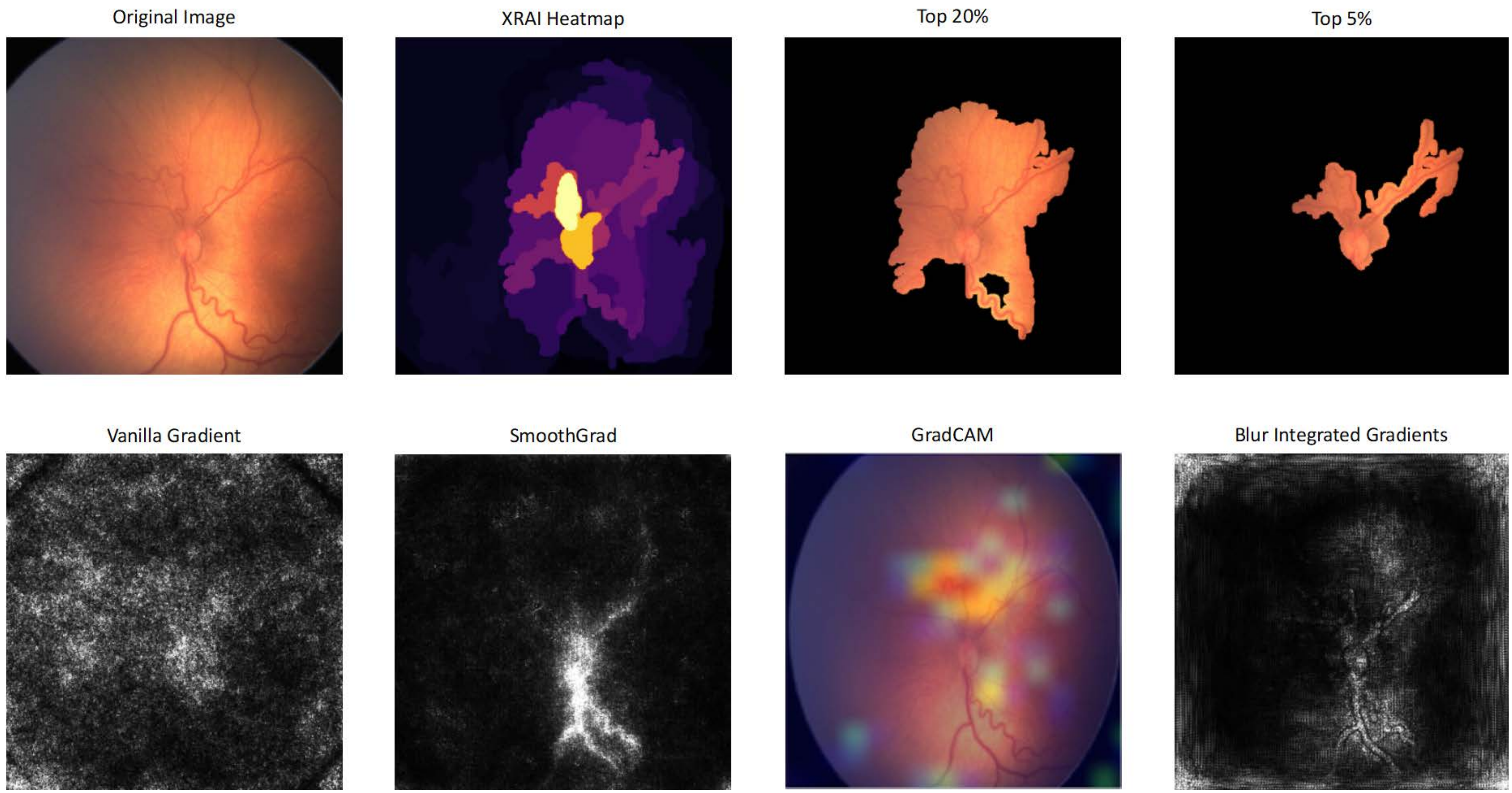
Supplementary Figures



Supplementary Figure 1: Example cases of Retcam disc-centred fundus photographs within the development dataset of newborns with **A**: normal, **B**: pre-plus and **C**: plus disease.



Supplementary Figure 2: Six examples of misclassification on the internal test set.



Supplementary Figure 3: Saliency maps using a range of techniques on an output from the bespoke model for plus disease.

References

1. Homerton [Internet]. [cited 2022 Apr 19]. Available from: https://www.citypopulation.de/en/uk/london/wards/hackney/E05009376__homerton/

2. Ravelo J, Adams G, Husain S. Identification of treatment-warranted retinopathy of prematurity by neonatal nurse specialist. Arch Dis Child Fetal Neonatal Ed [Internet]. 2021 Aug 23; Available from: http://dx.doi.org/10.1136/archdischild-2021-322266

3. English indices of deprivation 2019 [Internet]. GOV.UK. [cited 2022 Apr 19]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

4. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012 Jun 28;9(7):676–82.

5. 04_27 Examining the Premature and Infant Retina [Internet]. [cited 2021 Sep 2]. Available from: http://portal.e-lfh.org.uk/Component/Details/506306

6. Retinopathy of Prematurity: Case-Based Training [Internet]. 2015 [cited 2021 Sep 2]. Available from: https://www.aao.org/interactive-tool/retinopathy-of-prematurity-case-based-training

7. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021 Nov;3(11):e745–50.

8. Xu S, Venugopalan S, Sundararajan M. Attribution in Scale and Space. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. IEEE; 2020. Available from: http://dx.doi.org/10.1109/cvpr42600.2020.00970

9. Kapishnikov A, Bolukbasi T, Viegas F, Terry M. XRAI: Better attributions through regions. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) [Internet]. IEEE; 2019. Available from: http://dx.doi.org/10.1109/iccv.2019.00505

10. Simonyan K, Vedaldi A, Zisserman A. Deep inside Convolutional Networks: Visualising image classification models and saliency maps [Internet]. arXiv [cs.CV]. 2013. Available from: http://arxiv.org/abs/1312.6034

11. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise [Internet]. arXiv [cs.LG]. 2017. Available from: http://arxiv.org/abs/1706.03825

12. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020 Feb;128(2):336–59.

13. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients [Internet]. Vol. 1, Psychological Methods. 1996. p. 30–46. Available from: http://dx.doi.org/10.1037/1082-989x.1.1.30

14. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012;8(1):23–34.

15. Sawa J, Morikawa T. Interrater reliability for multiple raters in clinical trials of ordinal scale. Drug Inf J. 2007 Sep;41(5):595–605.

16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63.