Towards Clinical CMB Detection:
A Generalisable Two-stage Deep Learning Approach

Anita Karsa[1], Richard Joules[1], Jane Waugh[1], Frederik Barkhof[2,3], Luigi Lorenzini[2], Robin Wolz[1]

1 IXICO, London UK | 2 Department of Radiology & Nuclear Medicine, Amsterdam UMC, Vrije Universiteit, the Netherlands | 3 Queen Square Institute of Neurology and UCL Hawkes Institute, University College London, UK

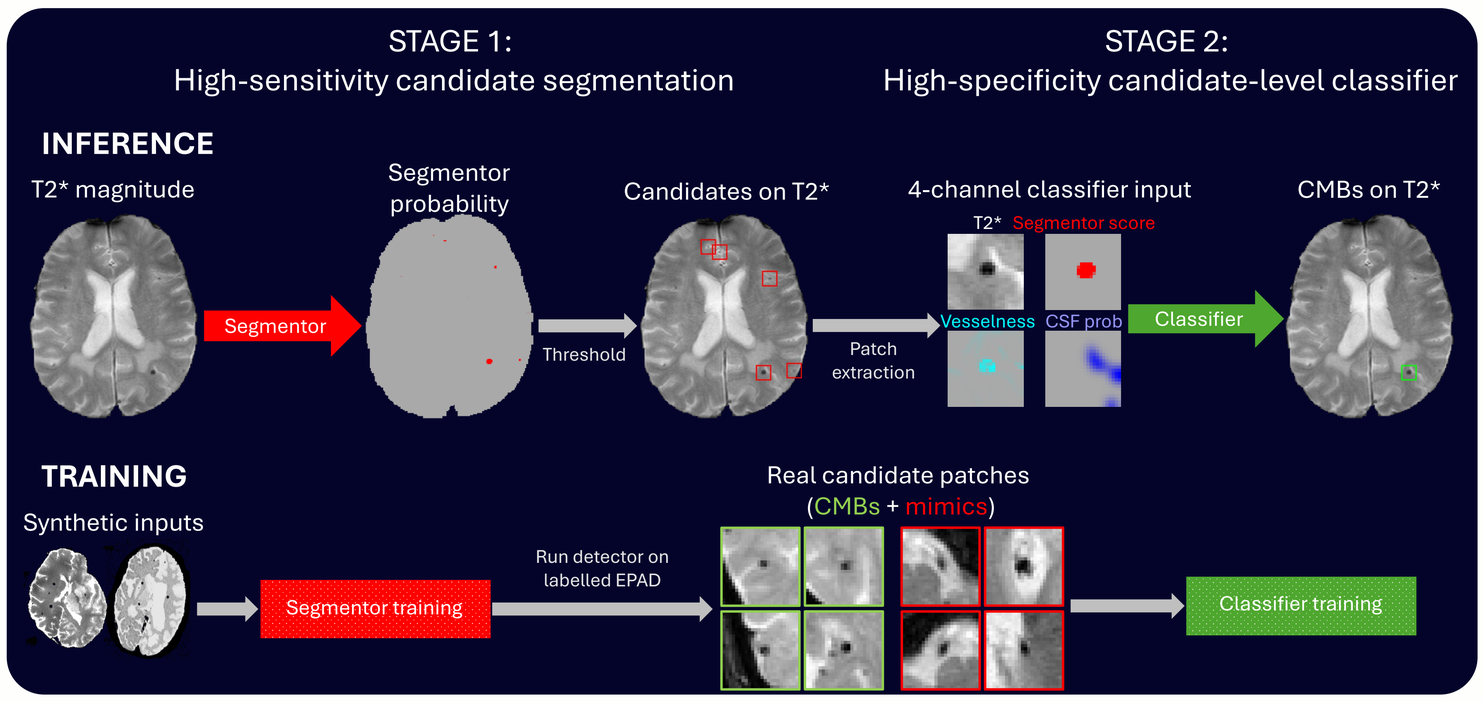
Key results

- We present a two-stage deep learning pipeline for cerebral microbleed (CMB) detection: a high-sensitivity segmentor trained on synthetic data, followed by a candidate-level classifier that separates true CMBs from mimics.
- At inference, model bagging (5-fold ensemble) and test-time augmentation (mirroring and rotation) make the predictions more robust.
- On 60 unseen, radiologist-labelled scans from two studies (ADNI, EPAD) and two sequences (T2*, SWI magnitude), **the pipeline achieved a mean subject-level sensitivity of around 90% at 1 false positive (FP) per subject**, exceeding the published performance of a commercially available FDA-cleared tool [1].
- Performance was consistent across datasets over the full range of classifier thresholds, so a single operating point can transfer across studies and sequences.
- **Next steps**: a larger real-world evaluation is being set up to test the pipeline as an assistive tool for radiologists in clinical reading workflows.

Background

Cerebral microbleeds (CMBs) are imaging markers of cerebral small vessel disease and are linked to higher risk of stroke, cognitive decline, and neurodegenerative pathology. CMB burden may also provide an informative marker of underlying cerebrovascular disease (CVD), helping to characterise vascular contributions to Alzheimer's disease (AD) progression and potential differences in treatment response and safety. In AD drug development, **CMBs are a key eligibility and safety measure**. Anti-amyloid monoclonal antibody therapies require close monitoring for treatment-related ARIA-H, including microhaemorrhages [2].

Manual radiological reads are labour-intensive [3] and subject to inter-observer variability, and ARIA-H monitoring at scale greatly increases this burden. **AI-based support to a human reader could speed up monitoring**, standardise reads across sites, and support both AD trials and the wider rollout of AD therapies. To be useful in multi-site trials, such tools must generalise across studies, scanners, and acquisition protocols.



Methods

Two-stage detection : An attention U-Net segmentor first identifies CMB-like candidates with high sensitivity. A ResNet classifier then separates true CMBs from false positives. Splitting the task lets each stage specialise: the segmentor on sensitive whole-brain detection, the classifier on fine-grained local differences between CMBs and mimics.

Training : The segmentor was trained on synthetic images with realistic simulated CMBs, generated on the fly to maximise variability and generalisation. The classifier uses four input channels:

- T2* magnitude
- Segmentor output
- A vesselness filter image, to flag elongated vessel-like structures
- A CSF probability map, to flag mimics in and around the CSF

The classifier was trained on real candidate patches (CMBs and mimics) created by applying the segmentor to radiologist-labelled EPAD T2* data [4] in a stratified, grouped 5-fold cross-validation design.

Inference : The outputs of all five cross-validation models were averaged. Test time augmentation was also applied to mirrored and rotated copies of each image, and those outputs were averaged also.

Evaluation : We evaluated on 60 unseen scans [4,5], each labelled by an expert radiologist:

- 20 EPAD T2* (spacing = 0.9 mm x 0.9 mm x 3 mm)
- 20 ADNI T2* (spacing = 0.8 mm x 0.8 mm x 4 mm)
- 20 ADNI SWI magnitude (T2* contrast at higher resolution, spacing = 0.5 mm x 0.5 mm x 1.8 mm)

Of the 60 subjects, 32 had mild cognitive impairment, 33 were amyloid-positive, and 21 and 11 were APOE ε4 heterozygous and homozygous reflecting a population relevant to anti-amyloid treatment and ARIA risk.

Operating points and assisted reading : Each candidate receives a classifier output score, so sensitivity and FP rate depend on the chosen threshold: lower thresholds favour sensitivity, higher thresholds favour specificity. The operating point can therefore be tuned to the application. In fully automated mode, we report sensitivity at 1 FP/subject. In an assisted-reading scenario, two thresholds split candidates into three groups. Candidates above an upper threshold (low FP) are accepted as CMBs, those below a lower threshold (low FN) are rejected, and only the uncertain candidates in between are sent for radiologist review.

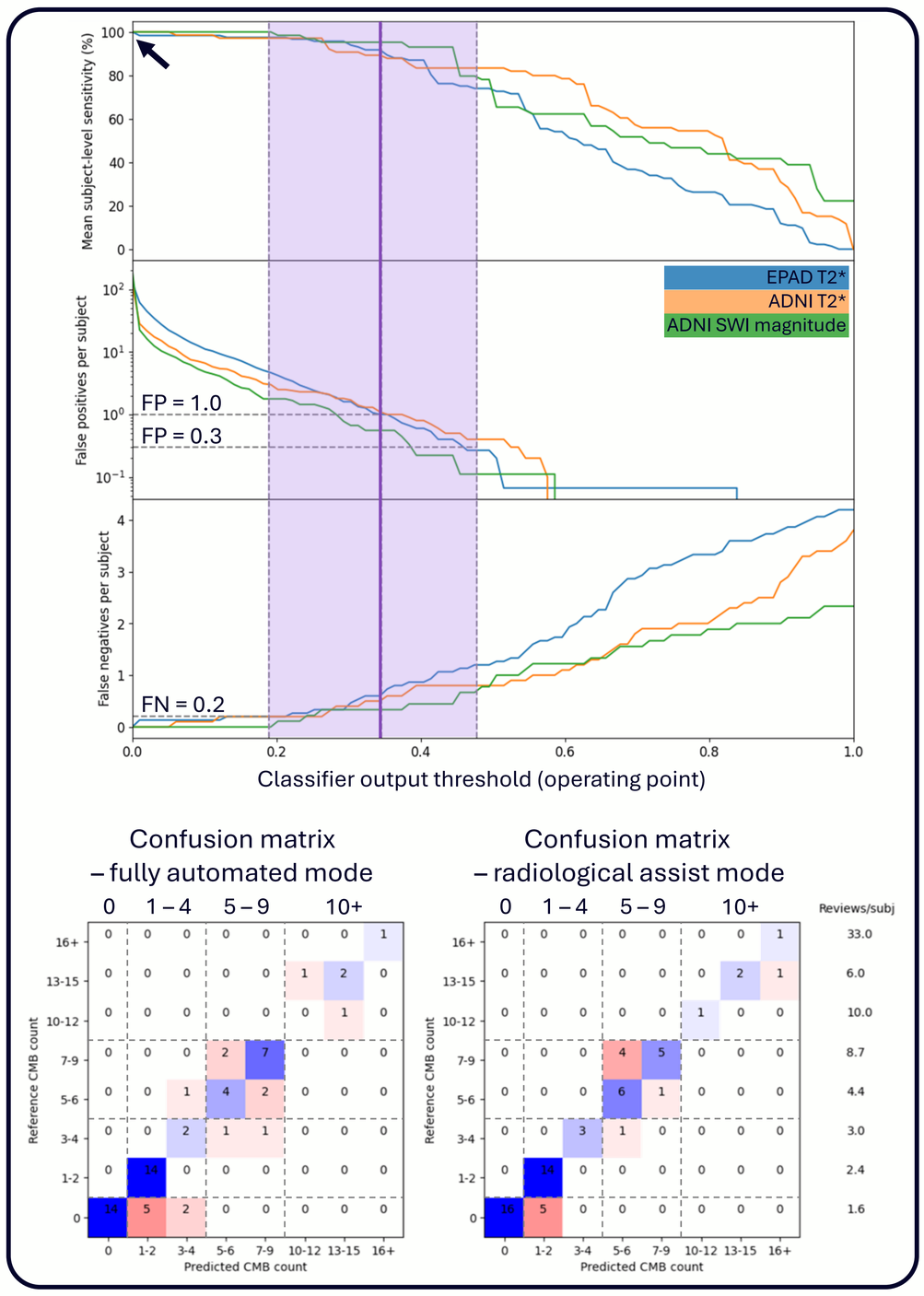
Disclosures : Anita Karsa, Richard Joules, Jane Waugh, and Robin Wolz are all employees of IXICO, a company developing commercial applications in this area. Frederik Barkhof is supported by the NIHR biomedical research centre at UCLH and is steering committee or Data Safety Monitoring Board member for Eisai and Idorsia, advisory board member for Combinostics, Scottish Brain Sciences, IXICO and Alzheimer Europe, consultant for Roche and Queen Square Analytics, and has research agreements with ADDI, GE Healthcare, and Roche.
Data acknowledgements : Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Data collection and sharing for ADNI is funded by the National Institute on Aging (NIH grant U19AG024904). Data were also obtained from the European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS; dataset [V.IMI], doi:10.34688/epadlcs_v.imi_20.10.30; ClinicalTrials.gov NCT02804789). The EPAD LCS was launched in 2015 as a public–private partnership led by Prof. Craig Ritchie, with the aim of providing a well-phenotyped population for modelling Alzheimer's disease in people without dementia. EPAD received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (grant agreement no. 115736) and an Alzheimer's Association grant (SG-21-818099-EPAD).

Results

Across all 60 external scans, the pipeline achieved a mean subject-level **sensitivity of around 90% at 1 FP/subject**, exceeding the published performance of a commercially available FDA-cleared tool [1]. Segmentor sensitivity was close to 100% for all three datasets (black arrow).

Sensitivity and FP rate followed closely matching curves across the three datasets over the operating range of classifier thresholds (purple range). This demonstrates how **one threshold can provide similar behaviour regardless of study or sequence**, which is important for deployment in multi-site trials without per-site calibration.

Performance on SWI magnitude images was comparable to T2*, despite the classifier being trained on T2* data only.



At an operating point of 1 FP/subject, the **fully automated pipeline correctly classified over 95% (20 of 21) of scans** with CMB counts in the ranges corresponding to moderate or severe ARIA-H (5–9 or 10+) [3]. The only error was a single near miss (assigned to an adjacent category), with no clinically serious misclassifications (left confusion matrix). **In an assisted-reading scenario**, where radiologists review only uncertain candidates, **accuracy in these categories rises to 100%** (right confusion matrix). Fewer than 30% of reviewed candidates changed label (positive to negative or vice versa) compared with the fully automated result.

By limiting expert review to a small number of uncertain candidates while maintaining consistent performance across studies and sequences, our tool offers a promising route to faster, more standardised CMB assessment in AD clinical trials.

References
[1] Sima, Diana, et al. *Neurology* 102.7_supplement_1 (2024).
[2] Hampel, Harald, et al. *Brain* 146.11 (2023): 4414-4424.
[3] Cogswell, Petrice M., et al. *American Journal of Neuroradiology* 43.9 (2022): E19-E35.
[4] Ritchie, Craig W., et al. *The Lancet Psychiatry* 3.2 (2016): 179-186.
[5] Jack, Clifford R., et al. *Journal of Magnetic Resonance Imaging* 27.4 (2008): 685-691.