Automated triaging of head MRI examinations using convolutional neural networks

David A. Wood[1], Sina Kafiabadi[2], Ayisha Al Busaidi[2], Emily Guilhem[2], Antanas Montvila[2], Siddharth Agarwal[1], Jeremy Lynch[2], Matthew Townend[3], Gareth J. Barker[4], Sebastian Ourselin[1], Thomas C. Booth[1,2], James H. Cole[4,5]

[1] King's College London, [2] King's College Hospital, [3] Wrightington, Wigan and Leigh NHSFT, [4] Institute of Psychiatry, Psychology and Neuroscience, King's College London,
[5] Dementia Research Center, University College London

Introduction

The growing demand for head magnetic resonance imaging (MRI) examinations, along with a global shortage of radiologists, has led to an increase in the time taken to report head MRI scans around the world [1]. For many neurological conditions, this delay can result in increased morbidity and mortality.

An automated triaging tool could reduce reporting times for abnormal examinations by identifying abnormalities at the time of imaging and prioritizing the reporting of these scans (Fig. 1). Convolutional neural networks (CNN) show considerable promise for this purpose, having achieved remarkable success on a range of medical imaging tasks[2][3]. However, a bottleneck to the development of a CNN-based tool for triaging routine hospital head MRI examinations is the difficulty of obtaining large, clinically-representative labelled datasets to enable supervised learning

In the last two years transformational developments within the field of NLP [4][5] have led to dramatic improvements in performance on a number of general as well as more specialised biomedical language tasks. As a result, it has recently become feasible to accurately automate the labelling of hospital head MRI examinations for computer vision applications [6]. The purpose of this study was to build on these breakthroughs and use a dedicated neuroradiology report classifier to generate a large labelled training dataset of MRI examinations from two large UK hospitals in order to train a deep learning-based abnormality detection model. We hypothesized that training at scale on clinically-representative data would result in generalizable models which are robust to variations in scanner vendors, imaging protocols and patient populations between different hospitals, and we sought to determine the generalisability of our models by training our models using different subsets of the available data (e.g. training on images from one hospital, testing on images from a second hospital etc.). We also we sought to quantify the clinical impact of using our model to triage out-patient head MRI examinations through a retrospective simulation study.

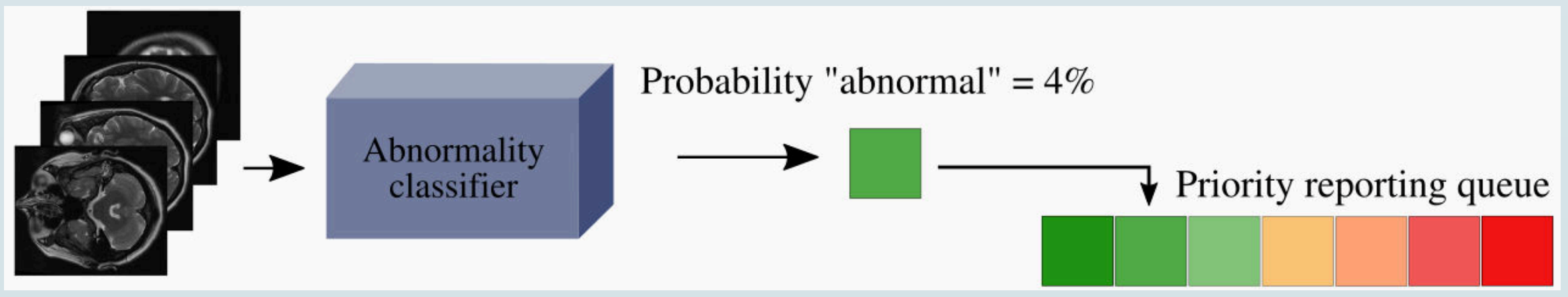

Figure 1: *Our classifier can be used to suggest the order in which head MRI examinations are reported by inserting images in real-time into a dynamic reporting queue based on the predicted likelihood of being abnormal (shown) or on the predicted category and time spent in the queue.*

Methodology

All 54,115 adult head MRI exams performed at King's College Hospital NHS Trust (KCH) and Guy's & St Thomas' NHS Trust (GSTT) between 2008-19 were obtained, along with the corresponding text reports produced by expert radiologists (UK consultant grade; US attending equivalent).

Using a validated deep learning-based neuroradiology report classifier [6][7], each examination was labelled as 'normal' or 'abnormal' (Fig. 2). Broadly speaking, findings which would generate a downstream clinical intervention were labelled 'abnormal', as were those which would be referred for case discussion at a multi-disciplinary team meeting.

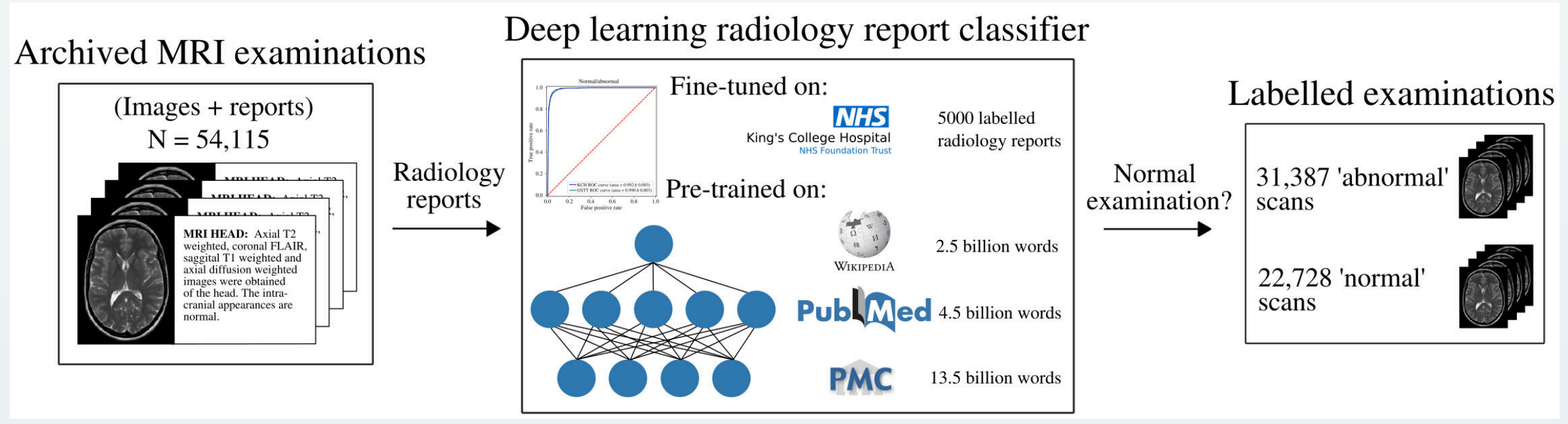
Archived MRI examinations
(Images + reports)
N = 54,115
MRI HEAD: Axial T2 weighted, coronal FLAIR, saggital T1 weighted and axial diffusion weighted images were obtained of the head. The intra-cranial appearances are normal.
Radiology reports
Deep learning radiology report classifier
Fine-tuned on:
NHS King's College Hospital NHS Foundation Trust
5000 labelled radiology reports
Pre-trained on:
Wikipedia
2.5 billion words
PubMed
4.5 billion words
PMC
13.5 billion words
Normal examination?
Labelled examinations
31,387 'abnormal' scans
22,728 'normal' scans

Figure 2: *Overview of dataset generation. All head 54,115 head MRI examinations (images and corresponding radiology reports) performed at KCH and GSTT between 2008 - 2019 were obtained. Using a dedicated deep learning-based neuroradiology report classifier, each examination was labelled as 'normal' or 'abnormal'.*

We trained (1) a baseline classification model, and (2) a classification model with an additional 'noise-correction' layer optimised for learning in the presence of label errors (Fig. 3). Both models utilize a 3D Densenet121 network for visual feature extraction, with the output of the final global average pooling layer concatenated with the patient's age and passed through a fully-connected layer (with softmax) to generate prediction probabilities for the two classes.

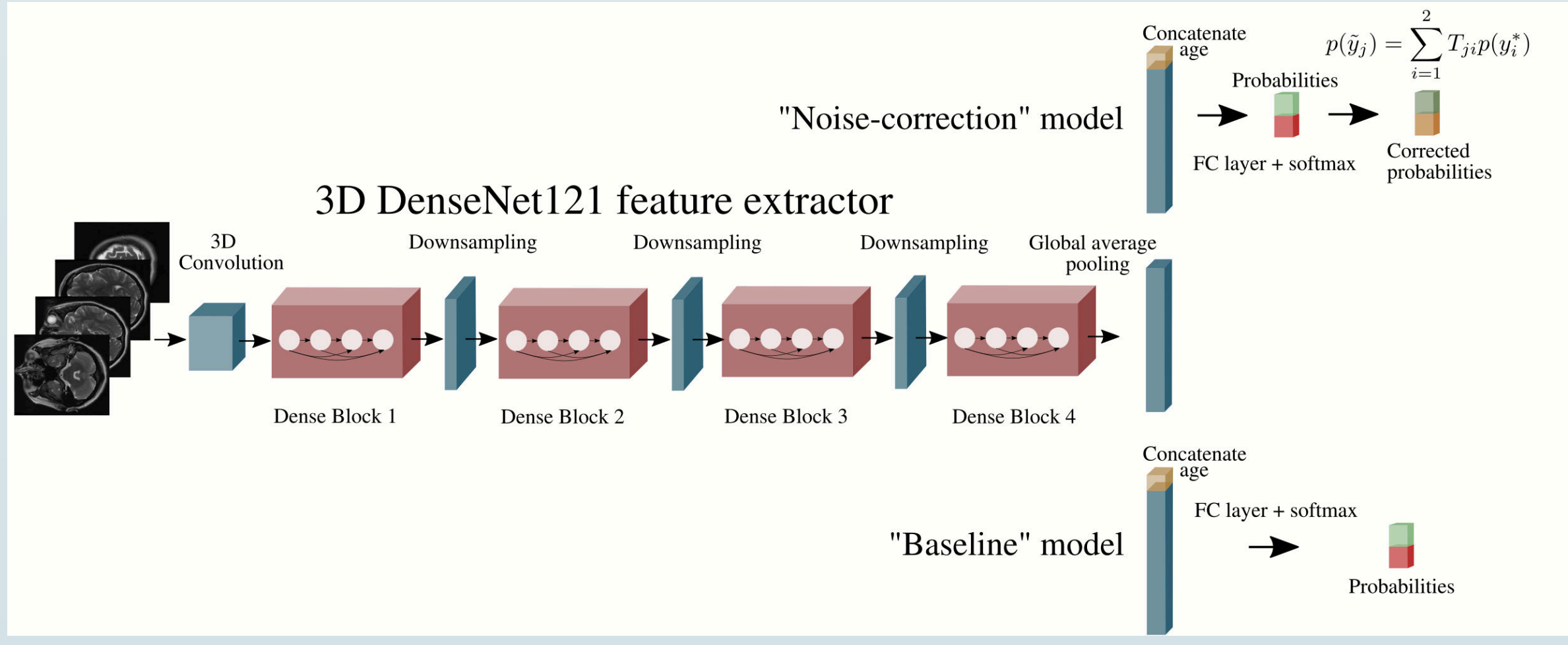

Figure 3: *Baseline classification model and 'noise correction' classification model. Both networks perform visual feature extraction using a 3D Densenet121, and concatenate this with the patient's age in order to generate class probabilities. The 'noise-correction' model includes an additional layer which modifies the predictions during training to enable learning the true, rather than the noisy, label distribution.*

To quantify the impact that our model would have in a real clinical setting, we performed a retrospective simulation study using all out-patient examinations performed at KCH and GSTT between 1/1/2018 - 31/12/2018 to determine what would have happened if our model had been used to suggest the order in which head MRI examinations were reported.

Results

Accurate classification (AUC > 0.9) was seen for both models for all training/testing combinations. However, 'noise-correction' led to a small but statistically significant improvement in all cases. When trained on scans from only a single hospital the models generalized to scans from the other hospital (ΔAUC ≤ 0.02) (Fig. 4, Table 1). Table 2 shows the impact that the best performing model (AUC = 0.943) would have had if it was used to suggest the order that examinations were reported. At both hospitals, the reduction in reporting times for abnormal examinations, as well as the increased reporting times for normal examinations, was statistically significant (p < 0.001

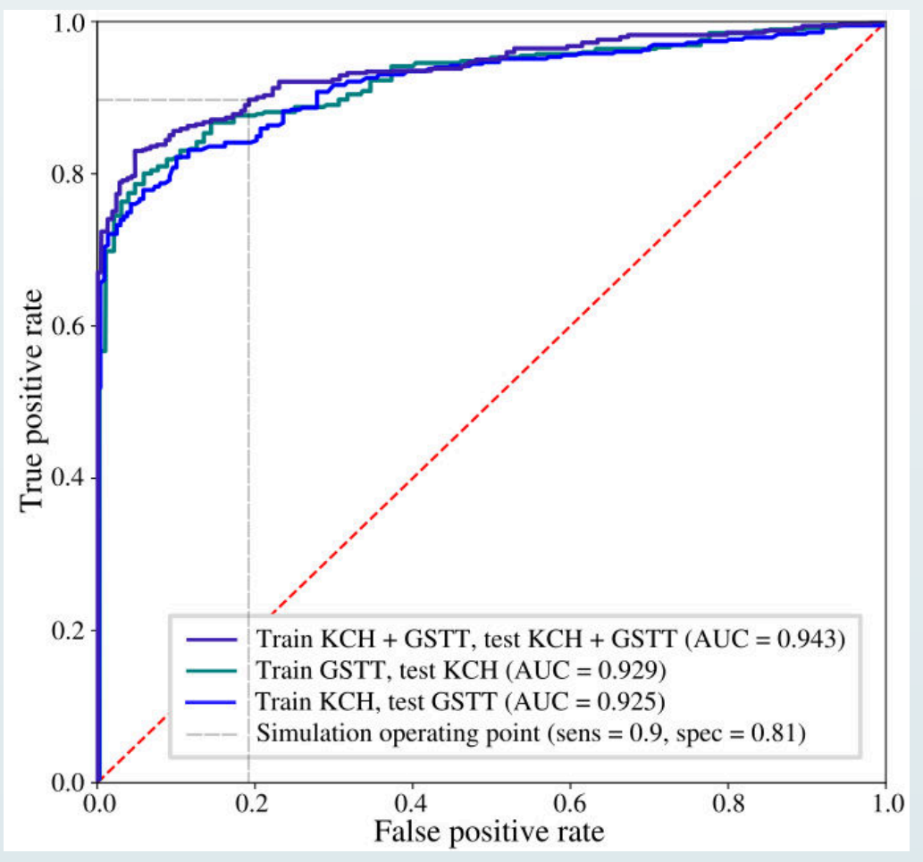

Figure 4: *Receiver operating characteristic curve for the 'noise-correction' model (1) trained/tested using images from both sites (purple), (2) trained on KCH, tested on GSTT (teal), and (3) trained on GSTT, tested on KCH (blue). Operating point used for the simulation study is also shown (dotted grey).*

Train		KCH			GSTT			Pooled		
Test		KCH	GSTT	Pooled	KCH	GSTT	Pooled	KCH	GSTT	Pooled
Model	Baseline	0.921	0.909	0.915	0.903	0.918	0.912	0.925	0.920	0.922
	Noise-corrected	0.941	0.925	0.933	0.929	0.931	0.930	0.946	0.939	0.943

Table 1: *Classification performance (AUC) for the baseline and 'noise-corrected' models. Both show accurate classification (AUC> 0.9), but 'noise correction' led to an improvement for all train/test splits (p < 0.05).*

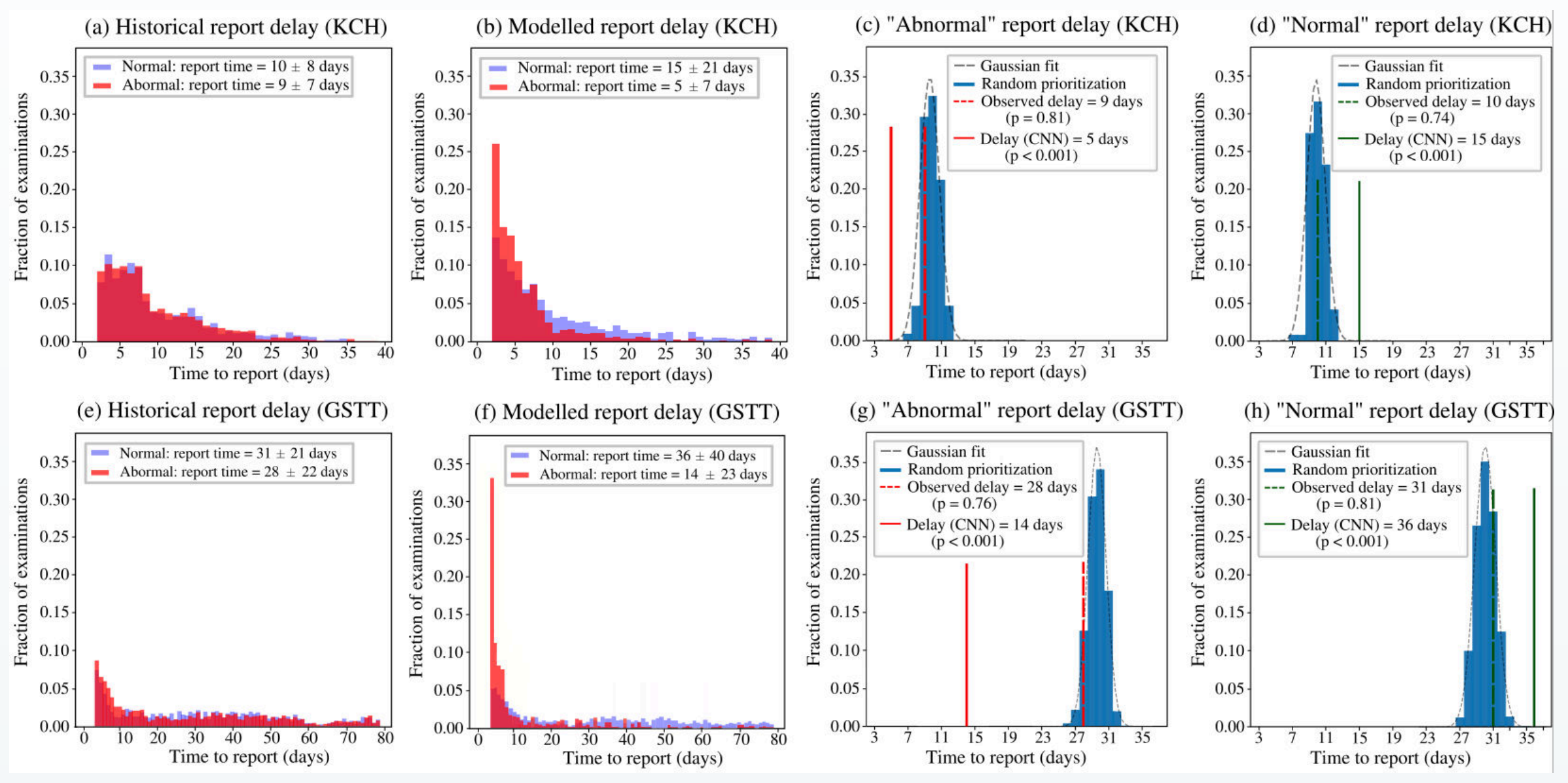

Figure 5: *Retrospective simulation results for KCH (top) and GSTT (bottom). Historical reporting delays (a, e) are compared with what would have been observed if our model had been used to prioritize the reporting of abnormal scans (b, f) at the two sites. To test for statistical significance, the null hypothesis distribution was generated (c, d, g, h) by repeating the simulation 1000 times, assigning a random priority to each examination (blue). At both sites, a statistically significant (p < 0.001) reduction in reporting times for abnormal examinations (solid red) compared with what was observed historically (dashed red) was seen.*

		Time to report	
		Normal	Abnormal
GSTT	Historical	31 ± 21 days	28 ± 22 days
	Our model	36 ± 40 days	14 ± 23 days
KCH	Historical	10 ± 8 days	9 ± 7 days
	Our model	15 ± 21 days	5 ± 7 days

Table 2: *Results of the retrospective simulation study, demonstrating the impact that our model would have on reporting times for abnormal scans at KCH and GSTT. Data are mean delay ± standard deviation.*

Conclusion

In this work we have presented a head abnormality classifier trained on 43,754 T2-weighted head MRI scans labelled using a neuroradiology report classifier, and demonstrated accurate classification on a test set of 800 scans containing over 90 classes of morphologically distinct abnormalities. We have shown that the model would reduce the time to report abnormal examinations at two UK hospitals, demonstrating feasibility as an automated triage tool.

Acknowledgements

This work was supported by the Royal College of Radiologists, King's College Hospital Research and Innovation, King's Health Partners Challenge Fund, and the Wellcome/Engineering and Physical Sciences Research Council Center for Medical Engineering (WT 203148/Z/16/Z)

References

[1] Claire E Bender, Swati Bansal, Darcy Wolfman, and Jay R Parikh. 2018 acr commission on human resources workforce survey. Journal of the American College of Radiology, 16 (4):508–512, 2019.
[2] Mauro Annarumma, Samuel J Withey, Robert J Bakewell, Emanuele Pesce, Vicky Goh, and Giovanni Montana. Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology, 291(1):196–202, 2019.
[3] Scott McKinney, et al.. International evaluation of an ai system for breast cancer screening. Nature, 577: 89–94, 01 2020. doi: 10.1038/s41586-019-1799-6.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[6] David A Wood, Jeremy Lynch, Sina Kafiabadi, et al. Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). In Medical Imaging with Deep Learning, pages 811–826. PMLR, 2020b.
[7] Wood et al., European Radiology 2021 (in press), Deep learning to automate the labelling of head MRI datasets for computer vision applications.