Creation and Testing of Model and Graphic User Interface

Datasets used are as in Table 1.

Table S1. Demographics of paediatric patients in the datasets used

Datasets	**Years covered by datasets**	**Number of patients**	**Male:Female ratio**	**Mean age (SD, range)**	**Total adhering to American College of Radiology guidelines (%)**
Training (split 80:20 for training and validation)	2006–2013	2470	1341:1129	8.2 (5.2, 0–20)	1887 (76%)
Test set	2014–2017	2711	1329:1382	8.4 (5.4, 0–20)	1661 (61%)
Audit set	2020	50	27:23	10.5 (5.0, 0–16)	20 (40%)

SD: standard deviation

Data labelling

Magnetic resonance imaging (MRI) orders in all datasets were manually labelled by research assistants under the supervision of an experienced paediatric radiologist and dichotomised based on whether each free-text order conformed to 2013 American College of Radiology (ACR) MRI brain guidelines, the reference standard. Text order entries that conformed to these guidelines were provided with a binary label: '1' if the order adhered to the guidelines and '0' if they did not conform. Labelled data were used for model training, validation and testing.

Model creation and evaluation

Data pre-processing for the free-text orders included standardisation of text to lowercase, correction of spelling errors, conversion to American spelling to be consistent with the style of ACR guidelines, and the removal of special characters and values to minimise the number of additional variables that could affect the prediction performance of the models.

Model creation utilised Python 3-based open source deep-learning library Keras (version 2.2.4). A stacked recurrent neural network configuration consisting of an ensemble of 3 bidirectional long short-term memory (Bi-LSTM) models in series was created and this works differently from the

conventional neural networks where filters are employed to transform data among the various layers of neurons across the network, as the series of Bi-LSTM models remember the input information and can enable textual patterns and sequences to be more accurately captured both in chronological and anti-chronological order.

Machine learning (ML) models trained using the bag-of-words methods, which are Random Forest, and the gradient boosted decision tree XGBoost, were also developed using the scikit-learn and XGBoost packages due to their widely reported success in ML studies, their robustness and high predictive performance. For the training of the ML models, greater data preprocessing is required as free-text data from the MRI orders had to be converted into structured variables that could be ingested by the bag-of-words models. Negation sentences such as “no headache” were removed as they were deemed a source of poor precision in medical entries. Words were crudely stemmed to their base form (such as “giddiness” and “giddy” to “giddi”) to reduce data dimensionality.

The outputs of the prediction models were binary classifications based on whether or not each MRI order complied with the ACR guidelines, using 50% as the cut-off. Standard evaluation metrics such as area under curve (AUC), accuracy, specificity, precision, recall, and F1 score were used to evaluate the models.

Performance of created models is shown in Table S2.

Table S2. Performance of the models created on the validation and test datasets

	Validation dataset (494 cases)			**Test dataset (2,711 cases)**		
Types of Models	Random Forest	XGBoost	Bi-LSTM	Random Forest	XGBoost	Bi-LSTM
Accuracy	0.915	0.899	0.917	0.807	0.789	0.825
AUC	0.959	0.950	0.960	0.889	0.864	0.892
Specificity	0.766	0.738	0.794	0.684	0.647	0.778
Precision	0.937	0.929	0.944	0.816	0.797	0.859
Recall	0.956	0.943	0.951	0.885	0.878	0.855
F1 score	0.946	0.936	0.947	0.849	0.836	0.857

AUC: area under curve; Bi-LSTM**:** bidirectional long short-term memory; MRI: magnetic resonance imaging

Graphic user interface

To ensure clarity of how the model derives its prediction, the Locally Interpretable Model-Agnostic Explainer (LIME) algorithm was used with the python package lime to create the model explainer. With the LIME explainer, each prediction outcome is explained by assigning individual words with feature weights based on their degree of agreement with ACR guideline adherence. Pyqt5 was used to develop a graphical user interface for ease of use.

Test implementation in the radiological workflow

To check model performance against current workflow processes, 50 paediatric MRI brain requests obtained in 2020 from an audit were used. Triage result from the best performing model was compared against those manually triaged by radiology staff with varying degrees of experience in MRI protocolling. Results from a senior paediatric radiologist were used as the reference gold standard. The other radiology staff involved were a junior paediatric radiologist, 9 radiology residents and 6 MRI radiographers. A non-medical research assistant inserted MRI orders into the graphic user interface to capture the percentage adherence to ACR guidelines with a percentage adherence less than 50% considered suitable for an ultrafast screening MRI brain protocol in the radiological workflow. Cohen's kappa coefficient was used to determine the inter-rater reliability between raters, where values closer to 1 meant higher agreement.

Table S3. Comparing classification of brain MRI orders by model and radiology staff

Designation	**Senior Paediatric Radiologist**	**Junior Paediatric Radiologist**	**3 Residents**	**6 Residents**	**6 MRI Radiographers**	**Bi-LSTM**
Years of radiology experience	17 years post-fellowship	10 years post-fellowship	0–2 years neuroradiology experience	3–4 years neuroradiology experience	3 to >10 years	Nil
Total ACR compliant orders	20	27	29	26	27	26
Kappa	Reference	0.72	0.42	0.68	0.72	0.67
P value	Nil	<0.01	<0.01	<0.01	<0.01	<0.01
Overcall	Reference	7	12	7	7	8
Undercall	Reference	0	3	1	0	0

ACR: American College of Radiology; Bi-LSTM**:** bidirectional long short-term memory; MRI: magnetic resonance imaging

Model for final deployment

In order to further improve the robustness of the model to be deployed, the Bi-LSTM was retrained and validated using the entire dataset (5,181 orders from 2006 to 2017) based on an 80:20 training and validation split. Performance of the final model is compared against other models used to analyse free text radiology reports in Table S4.

Table S4. Final deployed model compared against other models that have been used to analyse free text radiology reports

	MRI brain orders adherence to guidelines	**CT reports for presence of pulmonary embolism (Chen et al.[3])**		**MRI knee reports for injuries (Hussapour et al. [4])**		**MRI lumbar spine reports for type 1 Modic changes (Huhdanpaa et al. [5])**	**MRI brain reports for tumour status (stable, progressed, regressed) (Cheng et al. [6])**	**Radiology reports for critical findings (Lakhami et al. [7])**
Model	Bidirectional long short-term memory neural network (final deployed model)	Convolutional neural network		Support vector machine		Rule-based natural language processing	support vector machine	Rule-based natural language processing
Training and validation dataset (training: validation)	5181 (4:1)	2500 (2:3)		706 (4:1)	1748 (4:1)	200	541 (7:3)	2.3 million
Test dataset	Not applicable	1000 (internal)	859 (external)	1748	706	458	231	10 million
Accuracy	0.91	0.995	0.921	0.738	0.851		0.95 for stable tumour, 0.96 for progression, 0.94 for regression	
AUC	0.96	0.990	0.980					
Specificity	0.83	0.997	0.905			0.99	0.920	
Precision i.e. PPV	0.92			0.925	0.850	0.90	0.824	0.96
Recall i.e. Sensitivity	0.95	0.950	0.952	0.669	0.960	0.70	0.806	0.91
F1 Score	0.93	0.938	0.891	0.776	0.902	0.79	0.810	0.811.00

AUC: area under curve; CT: computed tomography; MRI: magnetic resonance imaging; PPV: positive predictive value

Superscript numbers: Refer to REFERENCES