DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Tools for Monitoring Effectiveness of AI Algorithms

Stuart R. Pomerantz
Massachusetts General Brigham Center for Clinical Data Science
MGH Neuroradiology
Harvard Medical School

Disclosures: None

Accuracy vs. Learning to Live with AI ‘warts & all’

- Dashboards & Interactive Analytics
- User Feedback
- Time Stamp/Report Time Analysis

Accuracy vs. Learning to Live with AI ‘warts & all’

- Dashboards & Interactive Analytics
- User Feedback
- Time Stamp/Report Time Analysis

Dashboards & Interactive Analytics: Dynamic Threshold Adjustment

- Adjust threshold based on post-deployment real-time performance
- Explore ability to adjust threshold for varying clinical scenarios
  - Worklist Prioritization vs Clinical-Decision-Support
  - High-Staffing vs Low-Staffing situations

ICH Accuracy Algorithm Background

nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-018-0324-9

An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets

Hyunkwang Lee[1,2,3], Sehyo Yune[1,3], Mohammad Mansouri[1], Myeongchan Kim[1], Shahein H. Tajmir[1], Claude E. Guerrier[1], Sarah A. Ebert[1], Stuart R. Pomerantz[1], Javier M. Romero[1], Shahmir Kamalian[1], Ramon G. Gonzalez[1], Michael H. Lev[1] and Synho Do[1*]

Owing to improvements in image recognition via deep learning, machine-learning algorithms could eventually be applied to automated medical diagnoses that can guide clinical decision-making. However, these algorithms remain a 'black box' in terms of how they generate the predictions from the input data. Also, high-performance deep learning requires large, high-quality training datasets. Here, we report the development of an understandable deep-learning system that detects acute intracranial haemorrhage (ICH) and classifies five ICH subtypes from unenhanced head computed-tomography scans. By using a dataset of only 904 cases for algorithm training, the system achieved a performance similar to that of expert radiologists in two independent test datasets containing 200 cases (sensitivity of 98% and specificity of 95%) and 196 cases (sensitivity of 92% and specificity of 95%). The system includes an attention map and a prediction basis retrieved from training data to enhance explainability, and an iterative process that mimics the workflow of radiologists. Our approach to algorithm development can facilitate the development of deep-learning systems for a variety of clinical applications and accelerate their adoption into clinical practice.

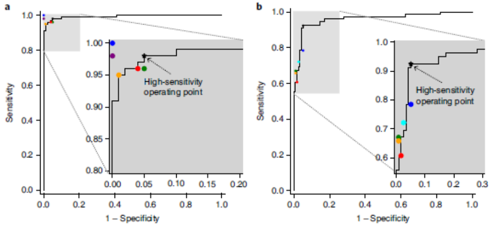


Fig. 4 | Test performance for ICH detection. ROC curves for ensemble model test performance (black lines) and radiologist performance (coloured circles) for the detection of ICH tested on the two separate datasets. **a**, The ensemble model tested with the retrospective dataset achieved an AUC value of 0.993 (95% CI of 0.982–0.999). Two radiologists outperformed the model, while three radiologists showed similar performance with that of the model. **b**, When tested with the prospective test dataset, the model achieved an AUC value of 0.961 (95% CI of 0.927–0.986) and showed higher sensitivity than any of the five radiologists at the predetermined operating point. Red, green and blue circles correspond to second-, third- and fourth-year radiology residents, respectively. Purple, cyan and orange circles correspond to attending radiologists with 9, 16 and 20 years of experience, respectively. ROC curves for each subtype of ICH are available in Supplementary Figs. 6 and 7.

Table 1 | Model performance on retrospective and prospective datasets in detecting and classifying ICH and its subtypes

	Retrospective			Prospective		
	AUC	Sensitivity (%)	Specificity (%)	AUC	Sensitivity (%)	Specificity (%)
ICH	0.993 (0.982, 0.999)	98.0 (95.3, 100)	95.0 (90.7, 99.3)	0.961 (0.927, 0.986)	92.4 (86.6, 98.2)	94.9 (90.9, 98.9)
IPH	0.980 (0.963, 0.993)	92.5 (85.4, 99.6)	91.8 (87.4, 96.2)	0.921 (0.843, 0.983)	68.8 (46.1, 91.5)	95.0 (91.8, 98.2)
IVH	0.979 (0.961, 0.992)	87.0 (78.0, 96.0)	95.9 (92.7, 99.1)	0.973 (0.910, 1.000)	83.3 (53.5, 100)	99.5 (98.5, 100)
SDH	0.959 (0.929, 0.983)	87.5 (77.3, 97.7)	86.9 (81.7, 92.1)	0.881 (0.812, 0.943)	70.5 (57.0, 84.0)	92.8 (88.7, 96.9)
EDH	0.922 (0.851, 0.978)	58.3 (30.4, 86.2)	95.2 (92.1, 98.3)	NA	NA	NA
SAH	0.960 (0.933, 0.980)	84.1 (75.7, 92.7)	88.5 (83.0, 94.0)	0.926 (0.883, 0.962)	76.3 (62.8, 89.8)	89.9 (85.2, 94.6)

The 95% CIs on the metrics are provided in parentheses. No cases of EDH were included in the prospective dataset.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Synho Do – Director, Laboratory of Medical Imaging and Computation

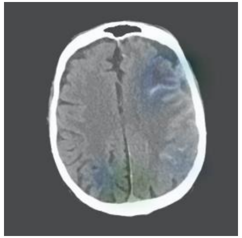
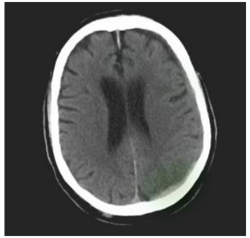
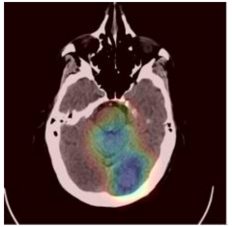
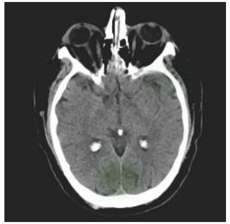
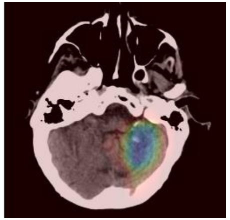
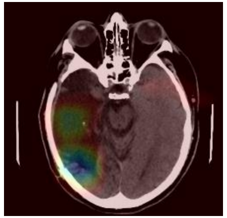
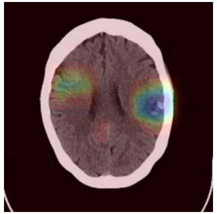
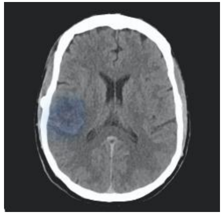
ICH Accuracy Algorithm Background

Do Model/PRIME 6-mo Implementation

ICH	Negative	3472						
	Positive	596						
	Total	4068						
Threshold	Sensitivity	Specificity	PPV	NPV	TN	FN	FP	TP
0.35	0.908	0.546	0.256	0.972	1896	55	1576	541
0.5	0.867	0.685	0.321	0.968	2378	79	1094	517
0.75	0.733	0.851	0.458	0.949	2955	159	517	437

Radiologist performance assessment. For comparison with the system, five radiologists with various levels of experience independently interpreted the two test datasets on the case level, based on the axial 5-mm series only, and blinded to clinical information and model output. The radiologists who interpreted the retrospective test dataset included first-, second- and third-year residents and two subspecialty board-certified neuroradiologists with 9 and 20 years of experience (radiologists B and C, respectively). For the prospective test, another board-certified neuroradiologist with 16 years of experience (radiologist E) replaced radiologist B, because radiologist B annotated the prospective dataset.

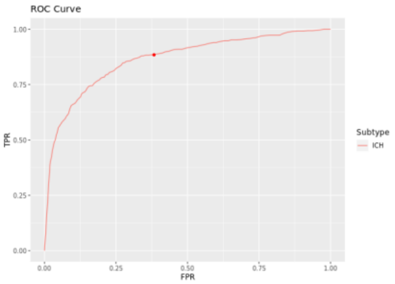
- No Clinical Context
  - Reason for Study
  - EHR review
  - Any potential discussion with referring clinicians/consults
- No thins sections
- No sag/coronal reformats
- No comparisons
  - Subsequent in-exam acquisitions (e.g. CTA/CTV series)
  - prior exams (e.g. stable falx dense thickening)
  - Simultaneously-acquired MRI (SWI) or immediate follow-up CT contemporaneous review

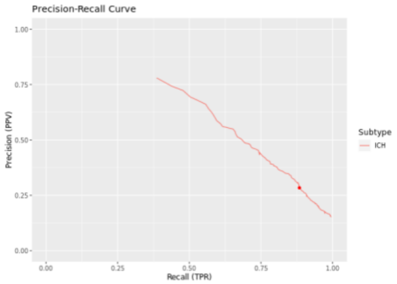
	Results description	Examples
True Positives	• True positives are not difficult cases for Radiologist to pick up	 
False Positives	• Hyper-dense Metastatic Disease • Peri-Hypodensity • Falx/ Tent • CP CA++ • Cavity and Craniectomy	    
False Negatives	• Analysis of prior and trauma indication led to positive ICH • Very small SDH • High threshold misses what may be an ”obvious” ICH (would be picked up at 50% threshold)	  

Dashboards & Interactive Analytics: Dynamic Threshold Adjustment

ICH Model Evaluation

Thresholds
ICH
0 — 0.42 — 1





Subtype	Sensitivity	Specificity	PPV	NPV
ICH	0.884	0.617	0.284	0.969

ICH - Metrics

```
          Reference
Prediction    N    P
         N 2143   69
         P 1329  527
```

```
Cases: 4068 Positive: 596 Negative: 3472
```

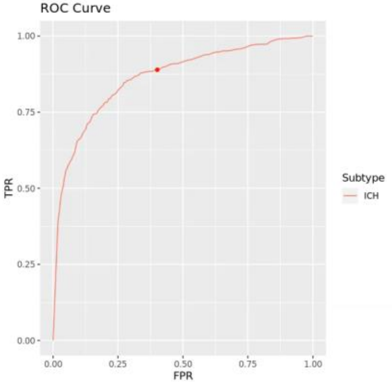
Show 10 entries   Search:

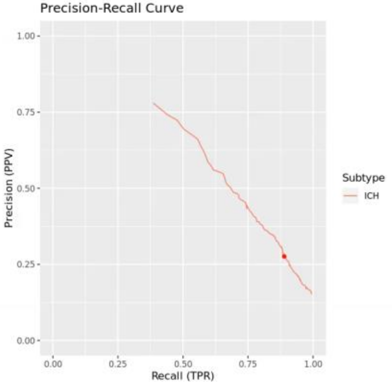
th	Sensitivity	Specificity	PPV	NPV
0	1	0		
0.01	1	0.023	0.149	1
0.02	0.995	0.046	0.152	0.981
0.03	0.993	0.063	0.154	0.982
0.04	0.993	0.078	0.156	0.985

Dashboards & Interactive Analytics: Dynamic Threshold Adjustment

ICH Model Evaluation

Thresholds
ICH
0 0.4 1
0 0.2 0.4 0.6 0.8 1





Subtype	Sensitivity	Specificity	PPV	NPV
ICH	0.889	0.599	0.276	0.969

ICH - Metrics

```
           Reference
Prediction    N    P
         N 2079   66
         P 1393  530
```

```
Cases: 4068 Positive: 596 Negative: 3472
```

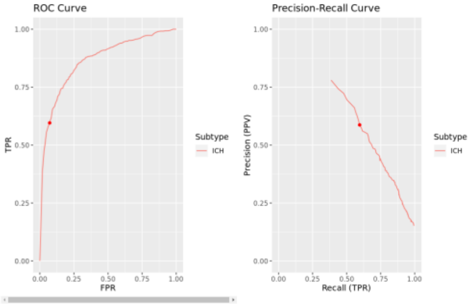
DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Dashboards & Interactive Analytics: Dynamic Threshold Adjustment

ICH Model Evaluation

Thresholds
ICH

0 — 0.9 — 1



Subtype	Sensitivity	Specificity	PPV	NPV
ICH	0.596	0.928	0.587	0.93

ICH - Metrics

0	1	0		
0.01	1	0.023	0.149	1
0.02	0.985	0.046	0.152	0.981
0.03	0.993	0.063	0.154	0.982
0.04	0.993	0.078	0.156	0.985
0.05	0.992	0.095	0.158	0.985
0.06	0.992	0.116	0.161	0.988
0.07	0.99	0.134	0.164	0.987
0.08	0.987	0.148	0.166	0.985
0.09	0.982	0.163	0.168	0.981

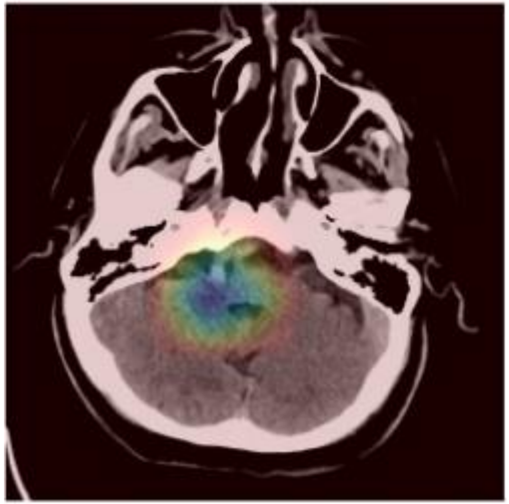
Showing 1 to 10 of 101 entries — Previous 1 2 3 4 5 … 11 Next

FP FN TP TN

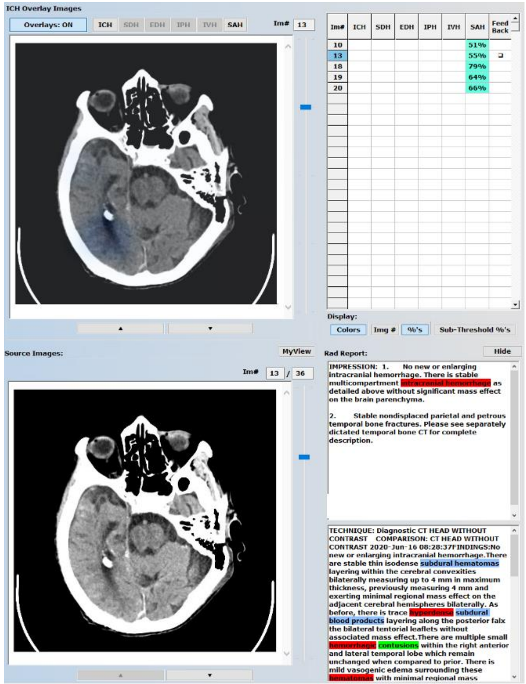
Show 10 entries — Search:

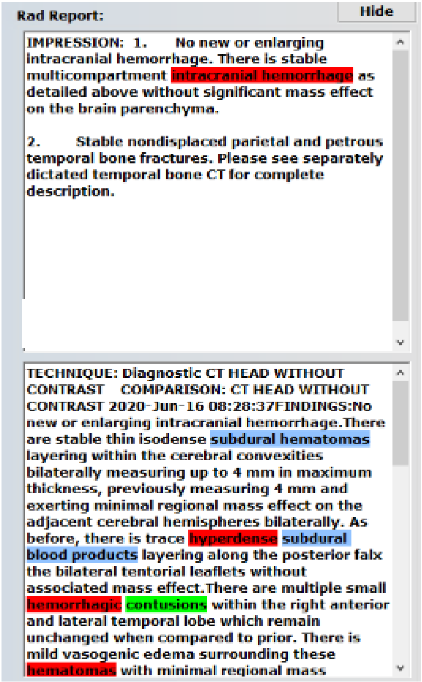
	StudyAccessionNumber	PatientID	Link
1	EDemo	PDemo	Images
2	EDemo	PDemo	Images
3	EDemo	PDemo	Images
4	EDemo	PDemo	Images
5	EDemo	PDemo	Images
6	EDemo	PDemo	Images
7	EDemo	PDemo	Images
8	EDemo	PDemo	Images
9	EDemo	PDemo	Images
10	EDemo	PDemo	Images

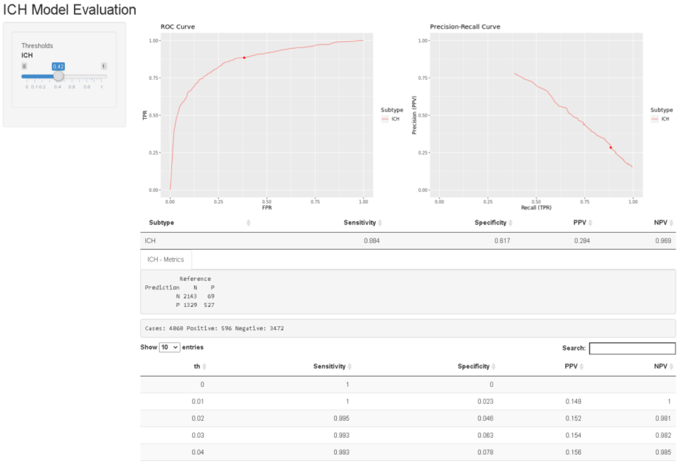
Showing 1 to 10 of 69 entries — Previous 1 2 3 4 5 6 7 Next

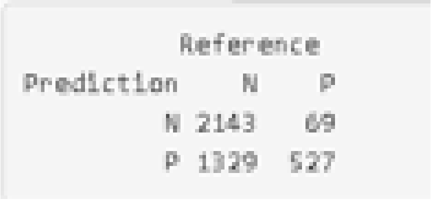


Report NLP - Automated Gold-Standard Generation









DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

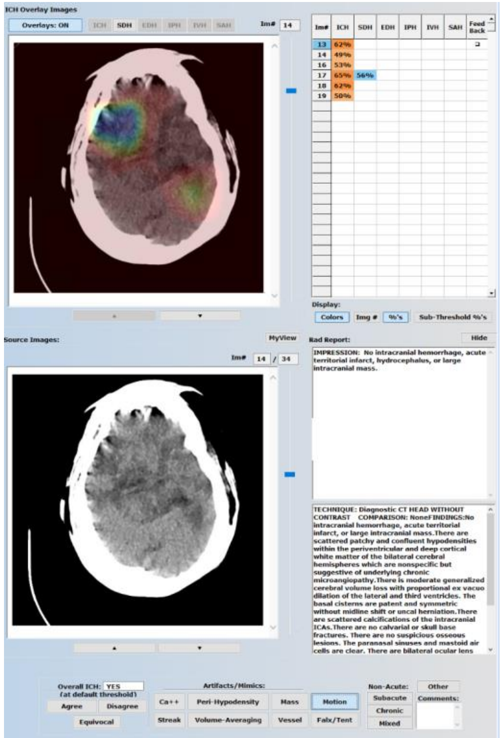
Dashboards & Interactive Analytics: Dynamic Threshold Adjustment

- Adjust threshold based on post-deployment real-time performance
- Explore ability to adjust threshold for varying clinical scenarios
  - Worklist Prioritization vs Clinical-Decision-Support
  - High-Staffing vs Low-Staffing situations

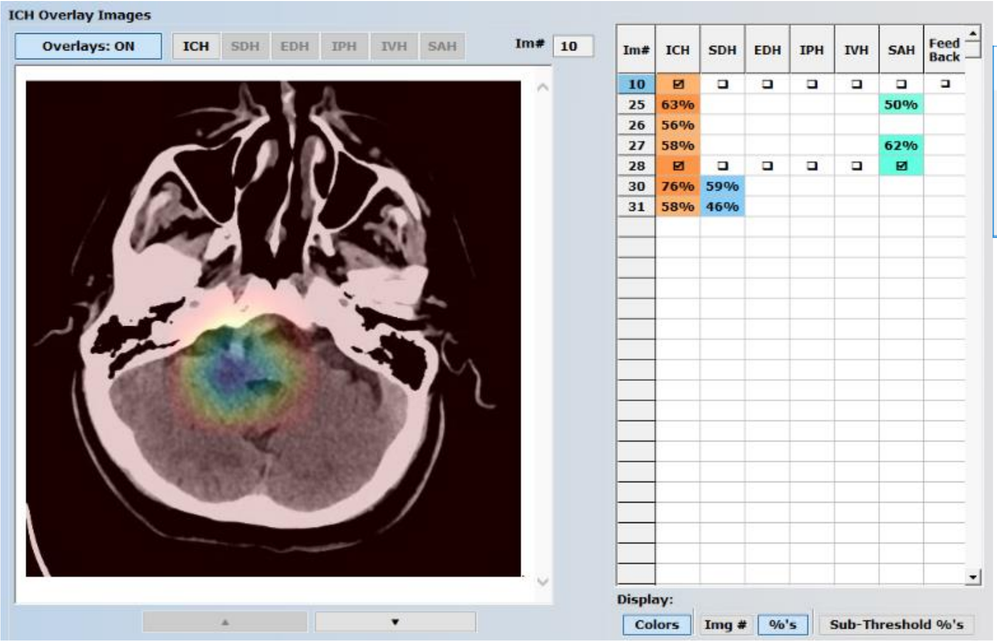
User Feedback Collection

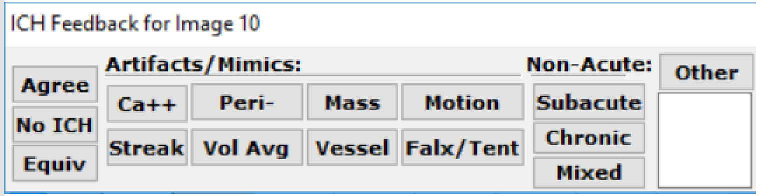
- Purpose
  - model refinement
  - Learn to Live with the algorithm – understand where users found it most helpful and where there are known pitfalls
- Concepts:
  - Thumb Up – Thumb Down insufficient
  - Model-Specific Feedback categories – known pitfalls
  - Whole Study Feedback
  - Granular Feedback

- Model-Specific Feedback categories – known pitfalls
- Whole Study Feedback

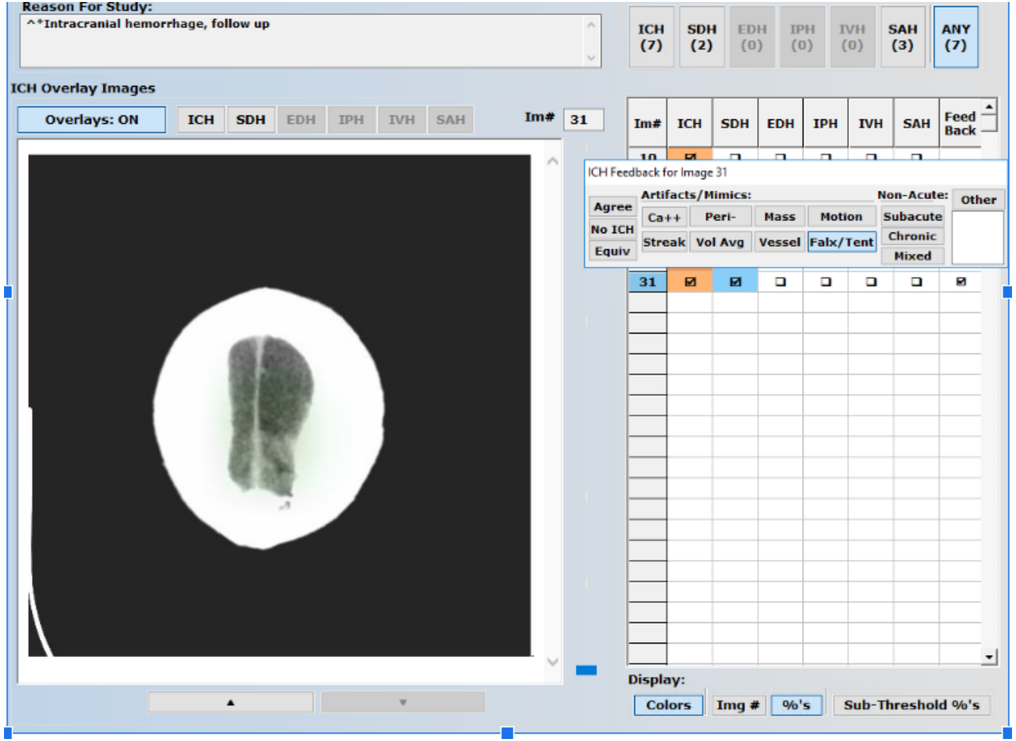


User Feedback Collection: Granular Feedback





User Feedback Collection: Granular Feedback

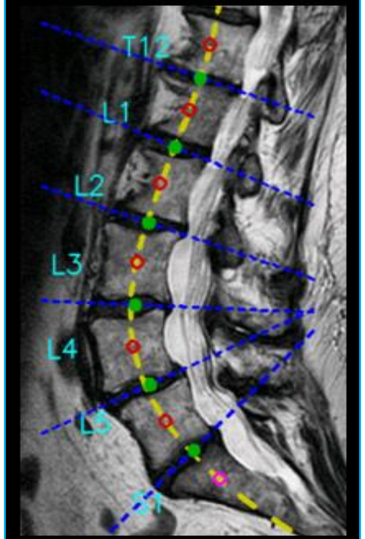


DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

Time-Stamp Assessment

- TAT: Prioritization of Urgent Studies
  - Pneumothorax
  - Intracranial Hemorrhage
- Reporting Time
  - Model Prediction of Disease Severity correlated with Interpretation Time

DeepSPINE: AI-Powered Diagnostic & Reporting Solution



Poly-fitting &
Disc Localization

SIGNIFICANT FINDINGS BY LEVEL:

T12-L1: There is mild spinal canal stenosis. There is no significant left foraminal stenosis and there is mild right foraminal stenosis.

L1-2: There is moderate spinal canal stenosis. There is mild left and right foraminal stenosis.

L2-3: There is severe central spinal canal and left and right foraminal stenosis.

L3-4: There is no significant spinal canal stenosis. There is moderate left foraminal stenosis and there is no significant right foraminal stenosis.

L4-5: There is mild central spinal canal and left and right foraminal stenosis.

L5-S1: There is mild spinal canal stenosis. There is mild left foraminal stenosis and there is moderate right foraminal stenosis.

IMPRESSION:

- **More efficient & accurate reporting**
- **More standardized grading & report descriptors**
- **Reduced interobserver variability**

Pomerantz/MGH/CCDS

Time Stamp/Report Time Analysis

- DeepSPINE Model-Generated Aggregate stenosis grading correlated with Interpretation Time
- Assess impact of utilization of DeepSPINE AI model on Reporting/TAT

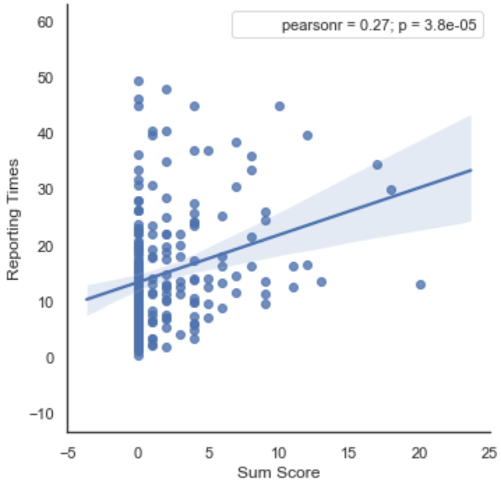
Report Create	Signed Final	Read Time
08:32:47	08:41:07	8m:20s
08:52:50	08:55:03	2m:13s
09:00:45	09:18:22	17m:37s
08:53:39	08:55:57	2m:18s
08:56:10	09:00:07	3m:57s

Stenosis Grades Extracted from:
- DeepSPINE Data Layer
- Rad-Generated Report Text

Grades Extracted from Report Text

T12-L1			L1-2			L2-3			L3-4			L4-5			L5-S1		
R	C	L	R	C	L	R	C	L	R	C	L	R	C	L	R	C	L
0	0	0	0	0	0	0	0	0	3	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0
0	0	0	0	0	0	1	1	1	0	0	0	3	0	0	3	0	1
0	0	0	0	0	0	0	0	1	0	3	0	2	3	2	0	1	0
0	0	0	0	0	0	1	1	0	1	0	1	1	0	1	1	0	1

Cumulative Severity Score
5
3
10
10
8



DeepSPINE: Smart Workflow Routing

DeepSPINE Data Layer

- Assess impact of utilization of DeepSPINE AI model on Reporting/TAT

- Predict Disease Severity/Interpretation Time
- Route to optimal staff/environment

- SIMPLE/SHORT
- MODERATE
- COMPLEX/LONG

MR LUMBAR SPINE - PRIORITIZED (51) 1, 1, 2, 2, 44,

+	Modality	Images	Study Description
	MR	250	MRI LUMBAR SPINE (BONE) WITHOUT CONTRAST
	MR	237	MRI LUMBAR SPINE (BONE) WITHOUT CONTRAST
	MR	173	MRI LUMBAR SPINE (NEURO) WITHOUT CONTRAST
	MR	265	MRI LUMBAR SPINE (NEURO) WITHOUT CONTRAST
	MR	191	MRI LUMBAR SPINE (NEURO) WITH AND WITHOUT
	MR	122	MRI LUMBAR SPINE (NEURO) WITHOUT CONTRAST
	MR	257	MRI LUMBAR SPINE (NEURO) WITH AND WITHOUT
	MR	119	MRI LUMBAR SPINE (NEURO) WITHOUT CONTRAST
	MR	101	MRI LUMBAR SPINE (NEURO) WITHOUT CONTRAST

Pomerantz/MGH/CCDS

Accuracy vs. Learning to Live with AI ‘warts & all’

- Dashboards & Interactive Analytics
- User Feedback
- Time Stamp/Report Time Analysis

Acknowledgements

- James Brink – Chair of Radiology
- Keith Dreyer – Vice Chair for Informatics
- MGH Divison of Neuroradiology
  - -Chief, R. Gilberto Gonzalez
- MGH Divison of MSK Radiology
- Tom Schultz - EMI – IT Infrastructure

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

MGH & BWH CENTER FOR
CLINICAL DATA SCIENCE

- Keith Dreyer – Chief Data Science Officer
- Kathy Andriole – Director of Research
- Data Scientists: Jen-Tang Lu, Stefano Pedemonte, Chris Bridge
- Software Engineers: Sean Doyle, Mark Walters
- Clinical Innovation Fellow: Nir Neumark
- Research Fellows/Residents: Walter Wiggins, M. Travis Caton

Synho Do –Director, Laboratory of Medical Imaging and Computation