Interpreting Deep Learning model predictions using Shapley Values

Jay Shah[1,2], Catherine Chong[2,3], Todd Schwedt[2,3], Visar Berisha[1,2], Jing Li[5], Katherine Ross[4], Gina Dumkrieger[3], Jianwei Zhang[1], Nathan Gaw[5], Simona Nikolova[3], Teresa Wu[1,2]

[1]Arizona State University, [2]ASU-Mayo Center for Innovative Imaging,
[3]Department of Neurology, Mayo Clinic, Phoenix, [4]Phoenix VA Health Care System,
[5]Georgia Institute of Technology

Outline of the talk

- Need for Explainability in Deep Learning
  - Landscape of methods
- Shapley values for explanation
- Research Problem
  - Results from SHAP
- Next Steps

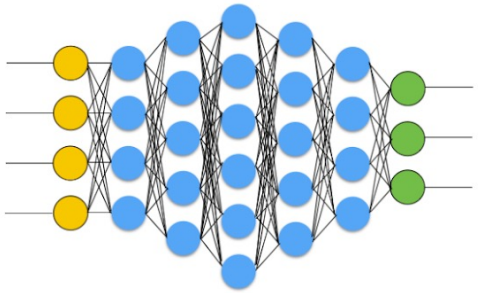
Outline of the talk

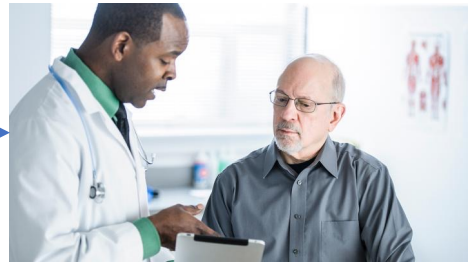
- **Need for Explainability in Deep Learning**
  - Landscape of methods
- Shapley values for explanation
- Research Problem
  - Results from SHAP
- Next Steps

Need for Interpretability in AI

- Understanding black-box models and their predictions are crucial in high stake applications

Is that it?



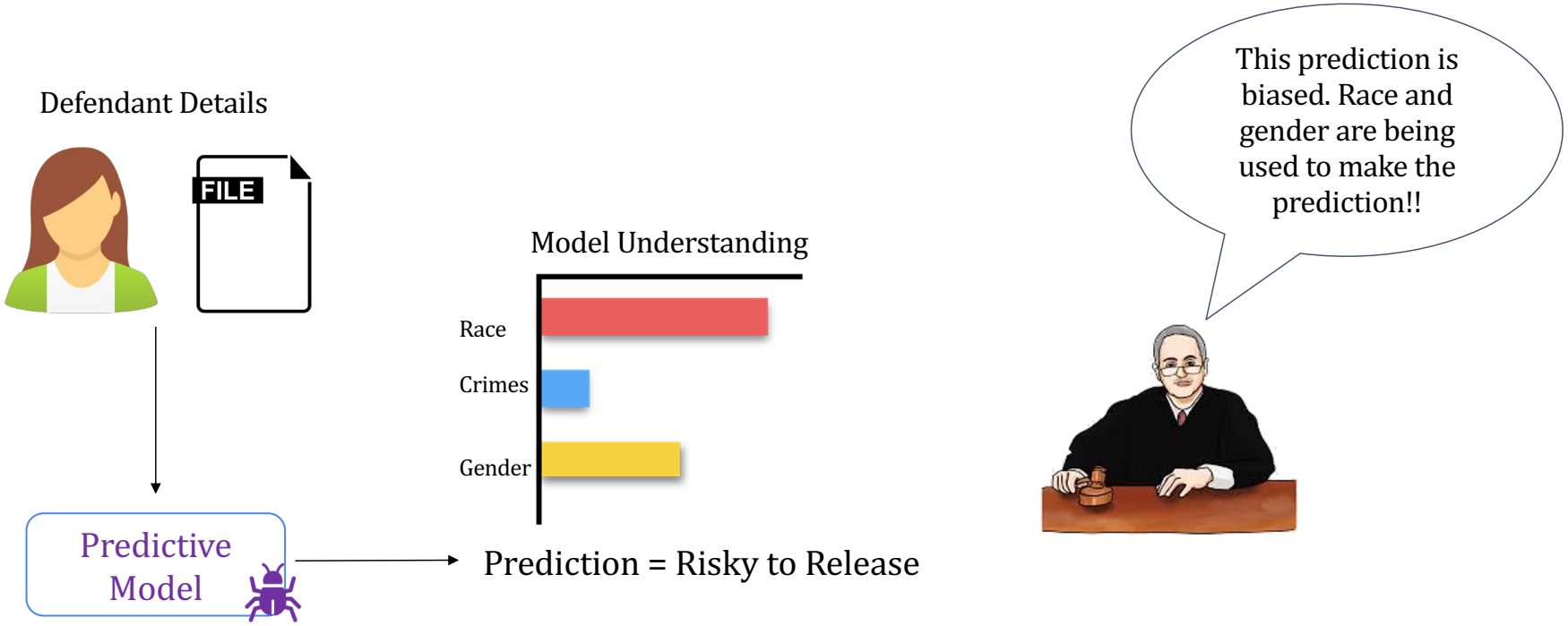


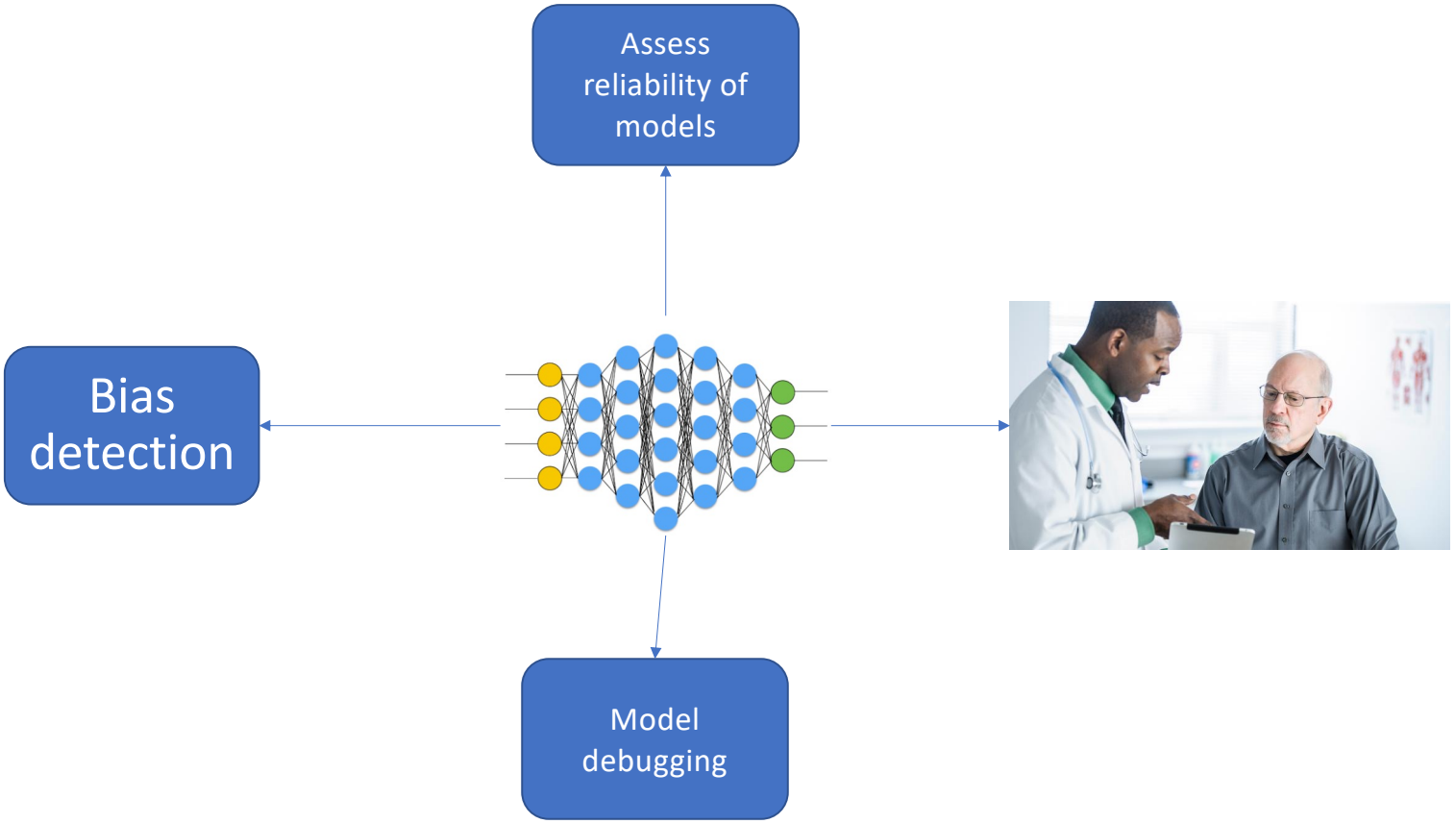
This example image, provided by the academics, of a cat has been modified so that when downscaled by an AI framework for training, it turns into a dog, thus muddying the training dataset

(above image)"Adversarial Preprocessing: Understanding and Preventing Image-Scaling Attacks in Machine Learning" [link]
"Backdooring and Poisoning Neural Networks with Image-Scaling Attacks" [link]

Is that it?

[Larson et. al. 2016]

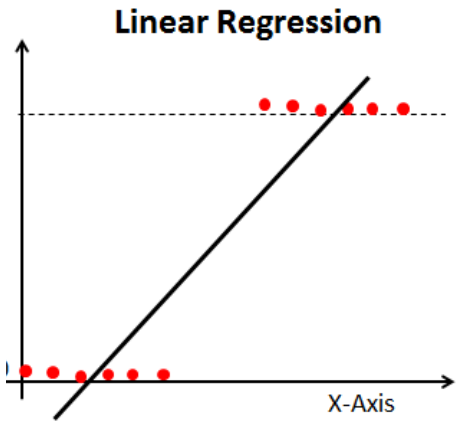


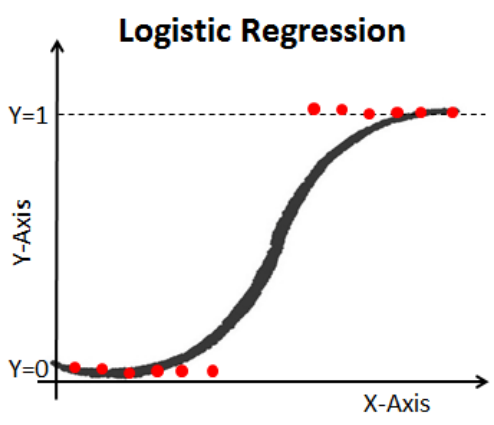

Assess reliability of models
Bias detection
Model debugging

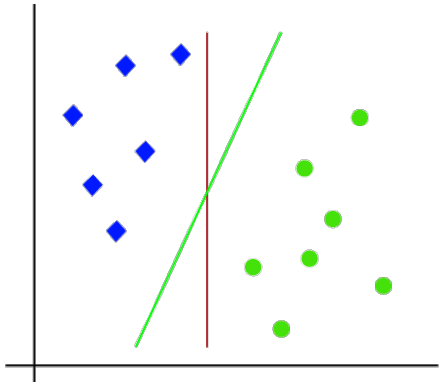
Towards model understanding

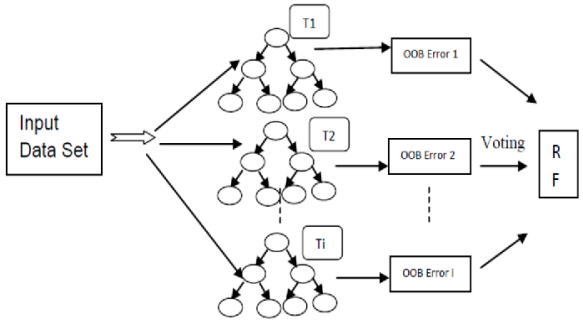
Approach 1:

Building inherently Interpretable models







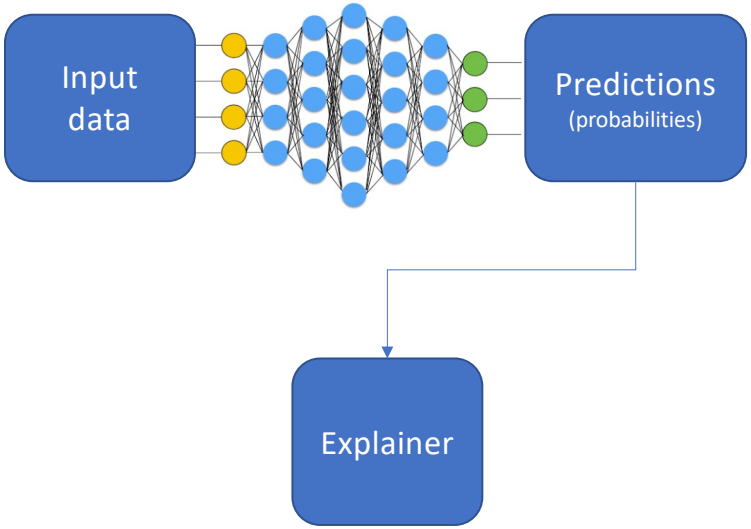


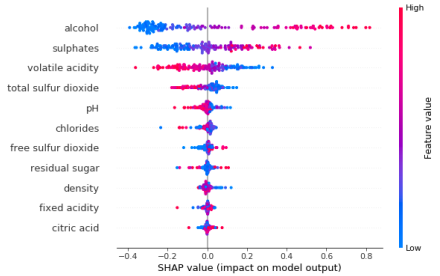
Towards model understanding

Approach 2:

Explaining black-box predictions using common human's understanding.

[post-hoc explainability]

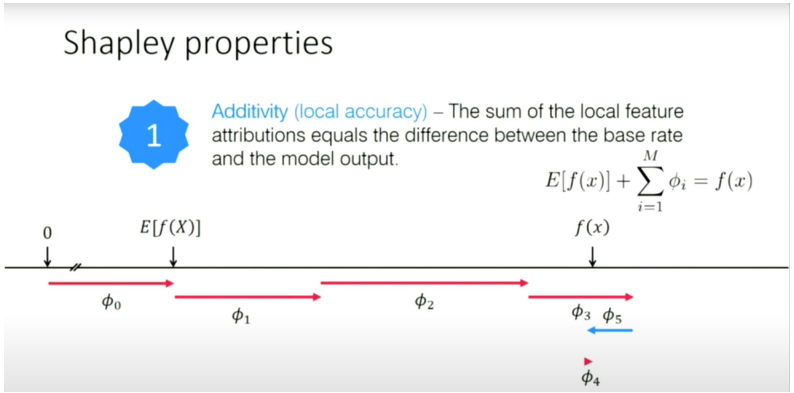


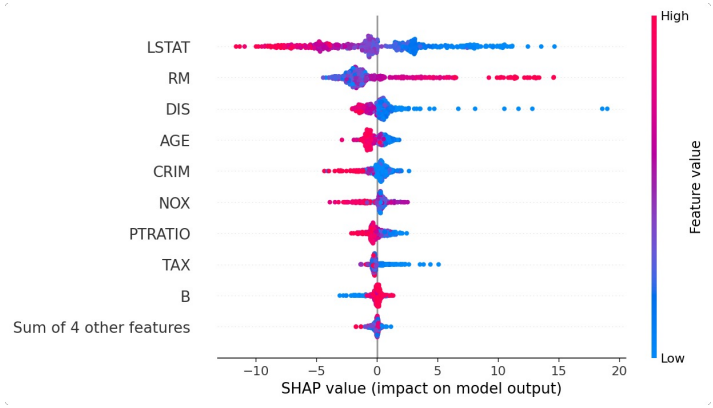


- **Local explanation based**
  - Feature importance
  - Counterfactuals
  - Instance based
  - Saliency maps

- **Global explanation based**
  - Concept-based
  - Proxy models/ Model distillation
  - Aggregating local explanations

- **Local explanation based**
  - Feature importance
  - Counterfactuals
  - Instance based
  - Saliency maps





Outline of the talk

- Need for Explainability in Deep Learning
  - Landscape of methods
- **Shapley values for explanation**
- Research Problem
  - Results from SHAP
- Next Steps

Intuition for Shapley Value

- A company has 2 employees, Jay and John

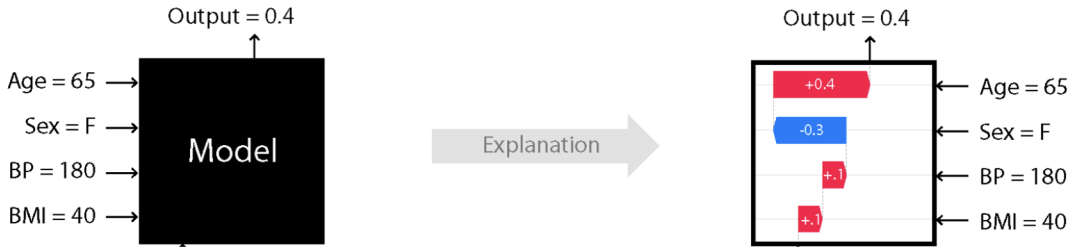
no one works, no profits	T{} = 0
Jay only works, company profit 20 units	T{Jay} = 20
John only works, company profit 10 units	T{John} = 10
Jay and John works, company profit 50 units	T{Jay, John} = 50

- How much does each deserve?

Permutation	**Marginal for Jay**	**Marginal for John**
Jay, John	Given Jay=20	we get John= (50-20) =30
John, Jay	we get Jay=(50-10) = 40	Given John=10
Shapley Value	30	20

Shapley value and Deep Learning

How are they connected?



output → risk probability for disease for that patient

Red → show features contributing positively to the disease classification

Blue → otherwise

Outline of the talk

- Need for Explainability in Deep Learning
  - Landscape of methods
- Shapley values for explanation
- **Research Problem**
  - Results from SHAP
- Next Steps

Post-Traumatic Headache (PTH)

- Two people can have same brain injuries yet only one might get PTH and other might not.
- It is a truly multi-variate study to understand the potential biomarkers and mechanisms for PTH.

- Goal: use DL for finding biomarkers across modalities; combined and separate

Post-Traumatic Headache (PTH)

- Two people can have same brain injuries yet only one might get PTH and other might not.
- It is a truly multi-variate s… potential biomarkers and…
- Goal: use DL for finding b… …and separate

MAYO CLINIC

Collaborating with Dr. Todd Schwedt & Dr. Catherine Chong, experts in neurology at the Mayo Clinic

→ Underlying biomarkers for PTH & Persistent-PTH. The goal is to build a multi-modal Deep Learning model to identify and delineate significant biomarkers.

Dataset(s)

NIH and DoD dataset of PTH and Migraine patients respectively

1. Clinical data
2. Imaging data
3. Speech (audio) data

1. Imaging Data

How the activation percentages per brain regions were generated?

Dataset

IXI: (https://brain-development.org/ixi-dataset/),
it has relatively larger cohort of HC (**N=581**), scans from three hospitals, including 1.5T (GE, Phillips) and 3T (Phillips), after age match, we have **N=451**
Age: 19.98 – 63.97 (mean: 43.13, SD=13.33)
Sex: 201 males/250 females

NIH:
N=58: 32 HC + 26 PTH
Age: 19-64 (mean: 39.12, SD=12.57)
Sex: 24/40 males/females

DOD:
N= 122: (35 MCM + 38 HC + 49 PTH)
Age: 19 – 63 (mean: 39.30, SD=10.63)
Sex: 61/61 males/females

3

Training and Visualization (+ Validation)

Overall workflow: we take ensemble approach (total 30 models, this workflow applies to each model)

01

PTH:
- 26 PTH (NIH)
- 49 PTH (DoD)

Migraine:
- 35 MCM (DoD)

HC:
- 32 HC (NIH)
- 38 HC (DoD)
- 451 HC (IXI)

02

Train until 100% AUC/Accuracy is reached

10 ResNet10,
10 ResNet18,
10 ResNet34

03

Visualize activations

Validation

4

Data Preparation Process (Continued)

How the activation percentages per brain regions were generated?

An example of raw Deep Learning activations

7

Results

From activations to brain locations

- Activations are in MNI space - [182 x 218 x182]
- Segmentation from Freesurfer

- Script
  - Conform FS outputs to RAS orientation
  - Impose and find top locations from Freesurfer labels

- Data: 26 PTH patients' activations from Deep Learning Model

8

Data Overview

Features = amount of activations in brain regions

Total brain regions = 176

Responses

Recovered -> 0

Not Recovered -> 1

% of activation from CNN

	A	B	C	D	E	F	G	H	FP	FQ	FR	FS	FT	FU	FV	FW
1	Patients	ctx-rh-parahippocampal	wm-rh-isthmuscingulate	ctx-rh-isthmuscingulate	Right-UnsegmentedWhiteMatter	ctx-rh-lingual	Left-Cerebellum-White-Matter	Brain-Stem	ctx-rh-superiorparietal	ctx-lh-superiorparietal	ctx-lh-inferiorparietal	ctx-lh-caudalmiddlefrontal	CC_Mid_Posterior	Right-vessel	Gina-Q	Gina-T
2	NIHPTH01	0.736738703	0.510493477	0.44378698	0.08527678	0.07765985	0.07617113	0.06424671	0	0	0	0	0	0	0	0
3	NIHPTH02	5.536831969	12.66331658	12.0910384	9.97166521	16.2509743	12.0889091	5.53961598	0.05610773	0.04842615	0.01584284	0.01460494	0	0	1	1
4	NIHPTH03	0.856531049	1.954674221	0.84779976	0.55142261	0.5407354	0.26987242	0.25762633	0	0	0	0	0	0	0	0
5	NIHPTH04	0	0	0	0.00186871	0	0	0	0.14227642	0	0	0	0	0	1	1
6	NIHPTH05	0	0	0	0	0.08835223	0	0	0	0	0	0	0	0	0	1
7	NIHPTH06	0	0	0	0.00888652	0	0	0.00926484	0	0	0	0	0	0	0	0
8	NIHPTH07	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
9	NIHPTH08	0	0.077459334	0.17094017	0.01397819	0.15974441	0	0.02075981	0.02053529	0	0	0	0	0	0	1
10	NIHPTH09	0	0	0	0	0	0	0	0	0	0	0.02200704	0	0	1	1
11	NIHPTH10	0.429184549	0.206568891	0	0.02489854	0.18849206	0.81507215	0.31751268	0	0	0	0	0	0	0	0
12	NIHPTH11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	NIHPTH12	0	0.094846665	0	0.27760732	1.07031121	0.87050408	0.35956852	0	0	0.03833254	0	0	0	0	1
14	NIHPTH13	0	0	0	0.00268219	0	0	0.05299928	0	0	0	0	0	0	0	0
15	NIHPTH14	0.862564692	0.189483657	0.55389859	0.8349801	0.41682456	0.25641026	0.40390644	0.06142506	0.01570352	0.02005817	0	0.34246575	0	0	1
16	NIHPTH15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	NIHPTH16	13.30913491	8.99589399	6.41784251	4.80898328	12.8706084	5.70926252	6.77269715	1.18464416	0.72274469	0.62671367	0.76558513	0	0	0	1
18	NIHPTH17	94.37819421	93.15068493	89.1304348	99.0816687	94.0915149	96.5334302	93.3998758	99.9759056	99.8201145	99.4684499	100	99.688958	100	1	1
19	NIHPTH18	0	0.202136876	0.19900498	0.07912018	0.12836971	0.02657807	0.11778029	0	0	0.21777004	0	0	0	0	1
20	NIHPTH19	0	0.174863388	0	0.05677126	0.13836478	0.00441306	0.06455986	0	0	0	0	0	0	0	0
21	NIHPTH20	0.247000706	0	0.02466091	0.72349257	2.22340994	0.0509684	0.26578073	1.54886132	0.50228311	0.83734561	2.01062216	0	0	0	1
22	NIHPTH21	0.237079184	0.02007226	0.37692085	0.36452994	1.16361909	0.01770147	0.00985675	0.40420372	0.04600345	0.14486124	3.47203817	0.32223416	0	1	1
23	NIHPTH22	0	0	0	0.02258042	0	0.00540073	0.00371927	0.00555247	0	0.04432624	0.11392117	0	0	0	0
24	NIHPTH23	0	0	0	0.26672462	0	0	0	0.00578871	0	0	0.73867367	0	0	0	0
25	NIHPTH24	34.53149002	59.52270621	61.258175	60.698349	49.3069307	67.2819118	53.7882319	51.9412382	51.2259444	49.4309262	53.8324873	52.4731183	16.6666667	0	0
26	NIHVAPTH01	0	0	0	0	0	0.02428953	0	0	0	0	0	0	0	1	1
27	NIHVAPTH02	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Data Overview

Features = amount of activations in brain regions
Total brain regions = 176

Objective
Find Correlation

Responses
Recovered -> 0
Not Recovered -> 1

% of activation from CNN

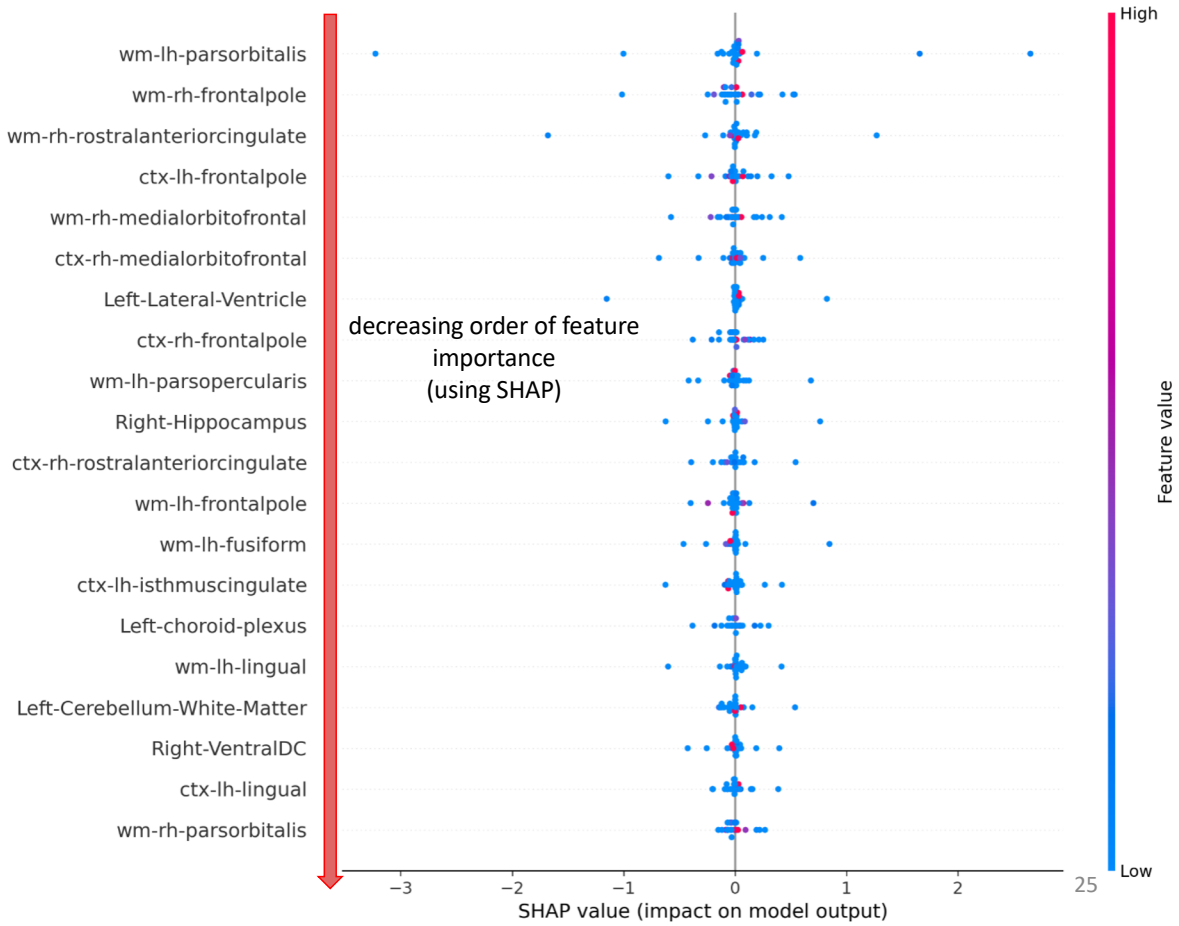
	A	B	C	D	E	F	G	H	FP	FQ	FR	FS	FT	FU	FV	FW
1	Patients	ctx-rh-parahippocampal	wm-rh-isthmuscingulate	ctx-rh-isthmuscingulate	Right-UnsegmentedWhiteMatter	ctx-rh-lingual	Left-Cerebellum-White-Matter	Brain-Stem	ctx-rh-superiorparietal	ctx-lh-superiorparietal	ctx-lh-inferiorparietal	ctx-lh-caudalmiddlefrontal	CC_Mid_Posterior	Right-vessel	Gina-Q	Gina-T
2	NIHPTH01	0.736738703	0.510493477	0.44378698	0.08527678	0.07765985	0.07617113	0.06424671	0	0	0	0	0	0	0	0
3	NIHPTH02	5.536831969	12.66331658	12.0910384	9.97166521	16.2509743	12.0889091	5.53961598	0.05610773	0.04842615	0.01584284	0.01460494	0	0	1	1
4	NIHPTH03	0.856531049	1.954674221	0.84779976	0.55142261	0.5407354	0.26987242	0.25762633	0	0	0	0	0	0	0	0
5	NIHPTH04	0	0	0	0.00186871	0	0	0	0.14227642	0	0	0	0	0	1	1
6	NIHPTH05	0	0	0	0	0.08835223	0	0	0	0	0	0	0	0	0	1
7	NIHPTH06	0	0	0	0.00888652	0	0	0.00926484	0	0	0	0	0	0	0	0
8	NIHPTH07	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
9	NIHPTH08	0	0.077459334	0.17094017	0.01397819	0.15974441	0	0.02075981	0.02053529	0	0	0	0	0	0	1
10	NIHPTH09	0	0	0	0	0	0	0	0	0	0	0.02200704	0	0	1	1
11	NIHPTH10	0.429184549	0.206568891	0	0.02489854	0.18849206	0.81507215	0.31751268	0	0	0	0	0	0	0	0
12	NIHPTH11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	NIHPTH12	0	0.094846665	0	0.27760732	1.07031121	0.87050408	0.35956852	0	0	0.03833254	0	0	0	0	1
14	NIHPTH13	0	0	0	0.00268219	0	0	0.05299928	0	0	0	0	0	0	0	0
15	NIHPTH14	0.862564692	0.189483657	0.55389859	0.8349801	0.41682456	0.25641026	0.40390644	0.06142506	0.01570352	0.02005817	0	0.34246575	0	0	1
16	NIHPTH15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	NIHPTH16	13.30913491	8.99589399	6.41784251	4.80898328	12.8706084	5.70926252	6.77269715	1.18464416	0.72274469	0.62671367	0.76558513	0	0	0	1
18	NIHPTH17	94.37819421	93.15068493	89.1304348	99.0816687	94.0915149	96.5334302	93.3998758	99.9759056	99.8201145	99.4684499	100	99.688958	100	1	1
19	NIHPTH18	0	0.202136876	0.19900498	0.07912018	0.12836971	0.02657807	0.11778029	0	0	0.21777004	0	0	0	0	1
20	NIHPTH19	0	0.174863388	0	0.05677126	0.13836478	0.00441306	0.06455986	0	0	0	0	0	0	0	0
21	NIHPTH20	0.247000706	0	0.02466091	0.72349257	2.22340994	0.0509684	0.26578073	1.54886132	0.50228311	0.83734561	2.01062216	0	0	0	1
22	NIHPTH21	0.237079184	0.02007226	0.37692085	0.36452994	1.16361909	0.01770147	0.00985675	0.40420372	0.04600345	0.14486124	3.47203817	0.32223416	0	1	1
23	NIHPTH22	0	0	0	0.02258042	0	0.00540073	0.00371927	0.00555247	0	0.04432624	0.11392117	0	0	0	0
24	NIHPTH23	0	0	0	0.26672462	0	0	0	0.00578871	0	0	0.73867367	0	0	0	0
25	NIHPTH24	34.53149002	59.52270621	61.258175	60.698349	49.3069307	67.2819118	53.7882319	51.9412382	51.2259444	49.4309262	53.8324873	52.4731183	16.6666667	0	0
26	NIHVAPTH01	0	0	0	0	0	0.02428953	0	0	0	0	0	0	0	1	1
27	NIHVAPTH02	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Feature importances from SHAP

using a
Logistic Regression model
(linear model)

⇩

Response Variable = Column T



Few observations

Based on some literature relating to Migraine, PPTH and PTH → we do find few overlapping regions as highlighted

Structural and Functional Brain Alterations in Post-traumatic Headache Attributed to Mild Traumatic Brain Injury: A Narrative Review

Todd J. Schwedt*

and anxiety scores, the PPTH group had significantly less cortical thickness in left and right frontal (superior frontal, caudal middle frontal, precentral) and right parietal (precuneus, supramarginal, inferior parietal, superior parietal) regions compared to healthy controls. Considering these regions that differed

among subjects with migraine and PPTH regarding brain regions related to pain processing, abnormally bilaterally reduced cortical thickness in frontal areas and right hemisphere parietal regions in PPTH subjects

→ The Relation between Persistent Post-Traumatic Headache and PTSD: Similarities and Possible Differences Martina Guglielmetti 1,2, Gianluca Serafini 3,4,* ,† , Mario Amore 3,4 and Paolo Martelletti 1,2,†

prefrontal cortex and anterior cingulate, has been observed in migraine patients

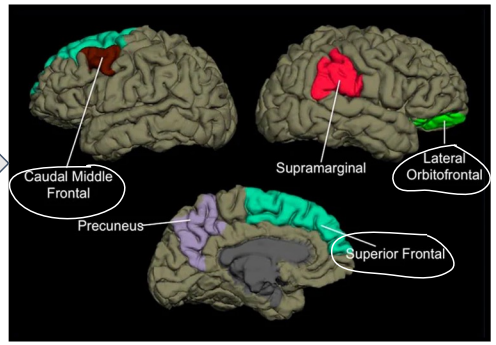
⟶ Altered functional magnetic resonance imaging resting-state connectivity in periaqueductal gray networks in migraine. Caterina Mainero MD, PhD, Jasmine Boshyan BS, Nouchine Hadjikhani MD, PhD

All patients	ctx-rh-superiorfrontal
	Right-UnsegmentedWhiteMatter
	Brain-Stem
	ctx-lh-rostralmiddlefrontal
	wm-rh-superiorfrontal
	wm-lh-rostralmiddlefrontal
	Right-Cerebellum-Cortex
	Left-Cerebellum-Cortex
	Left-UnsegmentedWhiteMatter
	ctx-lh-lateralorbitofrontal

Top regions sorted based on region-wise activations from DL model

Right-Cerebral-White-Matter
Left-Cerebral-White-Matter
ctx-lh-lateralorbitofrontal
ctx-rh-lateralorbitofrontal
ctx-lh-superiorfrontal
ctx-rh-superiorfrontal
ctx-lh-medialorbitofrontal
ctx-rh-medialorbitofrontal
ctx-rh-rostralmiddlefrontal
Right-Cerebellum-Cortex

Without region-wise activation sorting

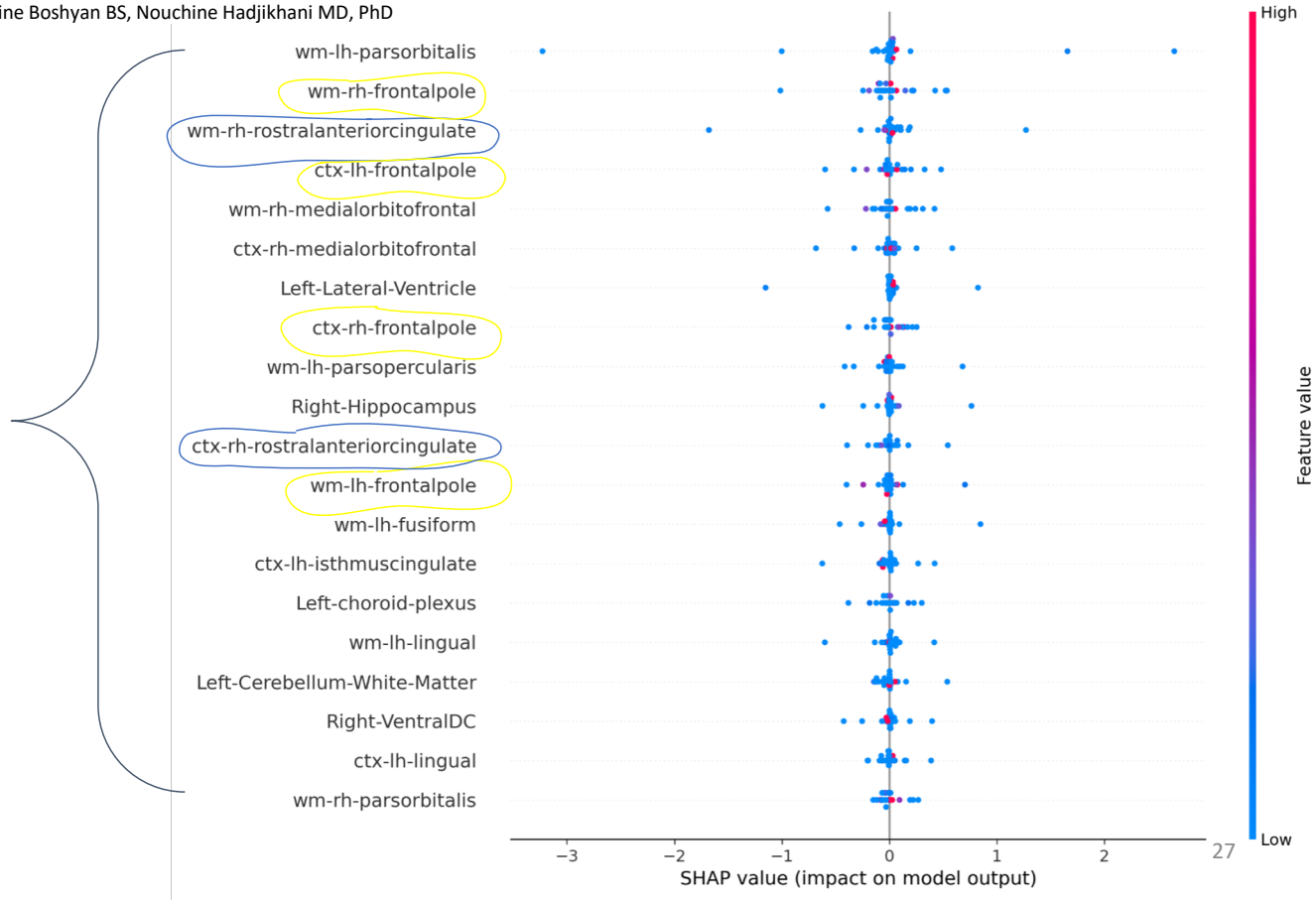


https://www.mayoclinic.org/medical-professionals/neurology-neurosurgery/news/post-traumatic-headache-clinical-trial-seeks-predictors-and-therapies/mac-20502096

prefrontal cortex and anterior cingulate, has been observed in migraine patients

→ **Altered functional magnetic resonance imaging resting-state connectivity in periaqueductal gray networks in migraine** Caterina Mainero MD, PhD, Jasmine Boshyan BS, Nouchine Hadjikhani MD, PhD

From SHAP's feature Importance list



2. Clinical data

Dataset Used

Clinical data only from the NIH study

Total sample size 64 patients

- 38 Healthy Controls (HC)
- 26 Post-Traumatic Headache (PTH)

- 790 features [age, hh_quality_1, midas_1, scat_pressure, ...]

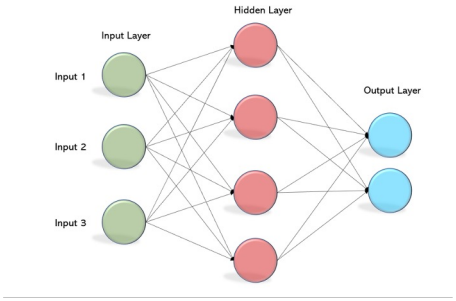
Questionnaire responses

2. Clinical data

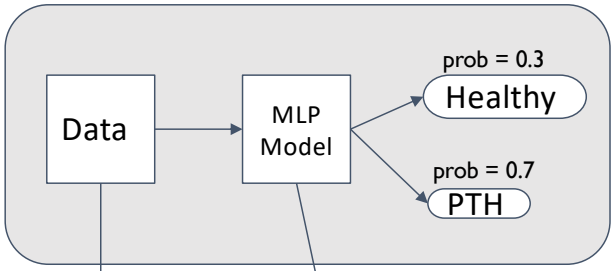
Model Used

- 2-layer MLP (Multi-Layer Perceptron) Network
  - 2-layer only because of simplistic clinical data

- Inputs are patient clinical features
- Outputs: HC or PTH



Interpretability Pipeline



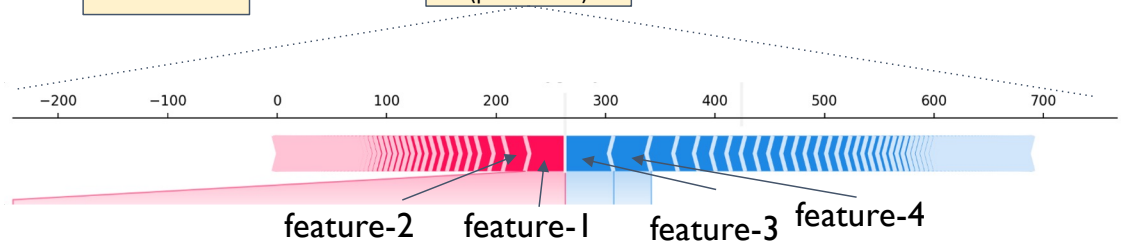
Standard classification pipeline

Goal is → Training + Explanation

patient #22

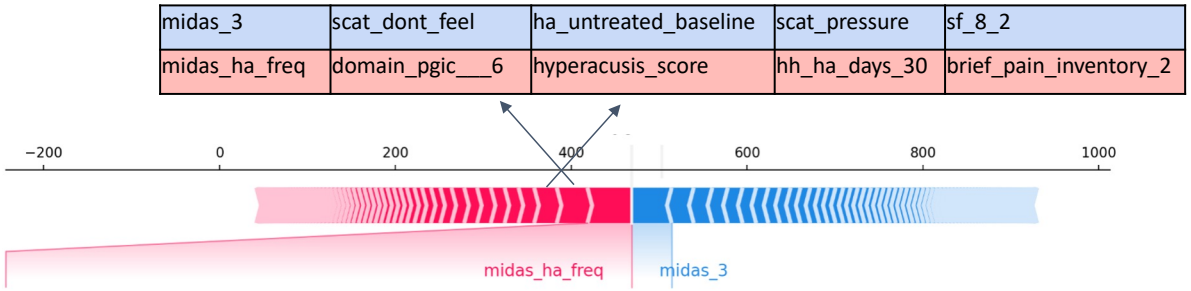
Explainer

Explainer Output (patient#22)



2. Clinical data

SHAP results



NIHHC01	asc_inter_score	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache
NIHHC02	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC03	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC04	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC05	hh_ha_days_30	midas_ha_freq	midas_3	paq_pho_sum	scat_headache
NIHHC06	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	age
NIHHC07	oiss_ohdas_points	oiss_walking_short	hh_ha_days_30	midas_ha_freq	oiss_standing_short
NIHHC08	midas_ha_freq	midas_3	scat_headache	hh_ha_days_30	scat_dont_feel
NIHHC09	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC10	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC11	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC12	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC13	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC14	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC15	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC16	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC17	midas_ha_freq	hh_ha_days_30	midas_3	scat_headache	ha_untreated_baseline
NIHHC18	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC19	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC20	hh_ha_days_30	scat_sad	midas_ha_freq	midas_3	scat_headache
NIHHC21	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC22	midas_ha_freq	hh_ha_days_30	midas_3	scat_headache	ha_untreated_baseline
NIHHC23	hh_ha_days_30	midas_ha_freq	midas_3	age	scat_headache
NIHHC24	oiss_ohdas_points	oiss_walking_short	oiss_standing_short	hh_ha_days_30	midas_ha_freq
NIHHC25	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC26	hh_ha_days_30	midas_3	midas_ha_freq	scat_headache	scat_dont_feel
NIHHC27	hh_ha_days_30	midas_ha_freq	scat_headache	midas_3	scat_dont_feel
NIHHC28	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	age
NIHHC29	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC30	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC31	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC32	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC33	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC34	midas_ha_freq	hh_ha_days_30	midas_3	asc_ictal_temp	scat_headache
NIHHC35	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC36	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	ha_untreated_baseline
NIHHC37	hh_ha_days_30	midas_ha_freq	midas_3	scat_headache	scat_dont_feel
NIHHC40	midas_ha_freq	hh_ha_days_30	midas_3	scat_headache	scat_dont_feel

Healthy cohort

major features:

hh_ha_days_30
midas_ha_freq
midas_3
scat_xxx

IIHPTH001	scat_dont_feel	bpi_pain_severity_2	age	bpi_pain_severity_1	scat_num_symp
IIHPTH002	hh_ha_days_30	midas_3	scat_dont_feel	midas_ha_freq	brief_pain_inventory_7
IIHPTH003	scat_headache	age	bpi_pain_severity_4	scat_pressure	scat_neck_pain
IIHPTH004	oiss_concentrating	oiss_weakness	bpi_pain_severity_2	oiss_vision	gad_2_2
IIHPTH005	hh_ha_days_30	midas_3	bpi_pain_severity_1	phq_2_1	domain_pgic___5
IIHPTH006	trails_b_zscore	domain_pgic___5	oiss_concentrating	hyperacusis_score	scat_hamen
IIHPTH007	hh_ha_days_30	midas_3	scat_headache	midas_ha_freq	asc_score_ic
IIHPTH008	scat_headache	midas_3	domain_pgic___5	bdi_score	scat_pressure
IIHPTH009	midas_3	hh_ha_days_30	scat_headache	scat_dont_feel	brief_pain_inventory_7
IIHPTH010	ha_untreated_baseline	hh_congestion	domain_pgic___1	sleep_scale_11	hh_conjunct
IIHPTH011	ha_untreated_baseline	scat_dont_feel	asc_score_ic	scat_headache	bpi_pain_severity_4
IIHPTH012	hh_ha_days_30	midas_ha_freq	scat_dont_feel	asc_score_ic	scat_headache
IIHPTH013	scat_headache	bdi_score	age	asc_score_ic	domain_pgic___5
IIHPTH014	bdi_score	domain_pgic___4	insomnia_total	oiss_ohdas_points	compass_total_score
IIHPTH015	domain_pgic___6	bdi_score	scat_neck_pain	scat_haphys	oiss_concentrating
IIHPTH016	bdi_score	age	insomnia_total	domain_pgic___7	hh_congestion
IIHPTH017	midas_ha_freq	midas_3	scat_headache	hh_ha_days_30	scat_dont_feel
IIHPTH018	midas_ha_freq	hh_ha_days_30	scat_headache	scat_dont_feel	nsi_headaches
IIHPTH019	domain_pgic___5	sleep_scale_10	bdi_score	scat_remembering	oiss_ohdas_points
IIHPTH020	phq_2_1	domain_pgic___6	age	oiss_ohdas_points	gad_2_2
IIHPTH021	midas_3	hh_ha_days_30	midas_ha_freq	ha_untreated_baseline	asc_score_ic
IIHPTH022	midas_ha_freq	domain_pgic___6	hyperacusis_score	hh_ha_days_30	brief_pain_inventory_2
IIHPTH023	midas_3	asc_score_ic	scat_dont_feel	hh_ha_days_30	midas_ha_freq
IIHPTH024	midas_3	domain_pgic___4	ha_untreated_baseline	hh_ha_days_30	bdi_score
IIHVAPTH01	hh_ha_days_30	midas_ha_freq	scat_headache	asc_score_ic	hh_days_abort
IIHVAPTH02	scat_headache	midas_ha_freq	hh_ha_days_30	asc_score_ic	oiss_concentrating

PTH cohort

major features:

hh_ha_days_30
midas_ha_freq
midas_3

scat_xxx

Observations

- For Healthy subjects,
  - the four important features are homogeneously distributed

- For PTH,
  - interestingly similar set of features,
  - way the features contributing to the diagnosis varies greatly → due to the heterogeneity of the patients

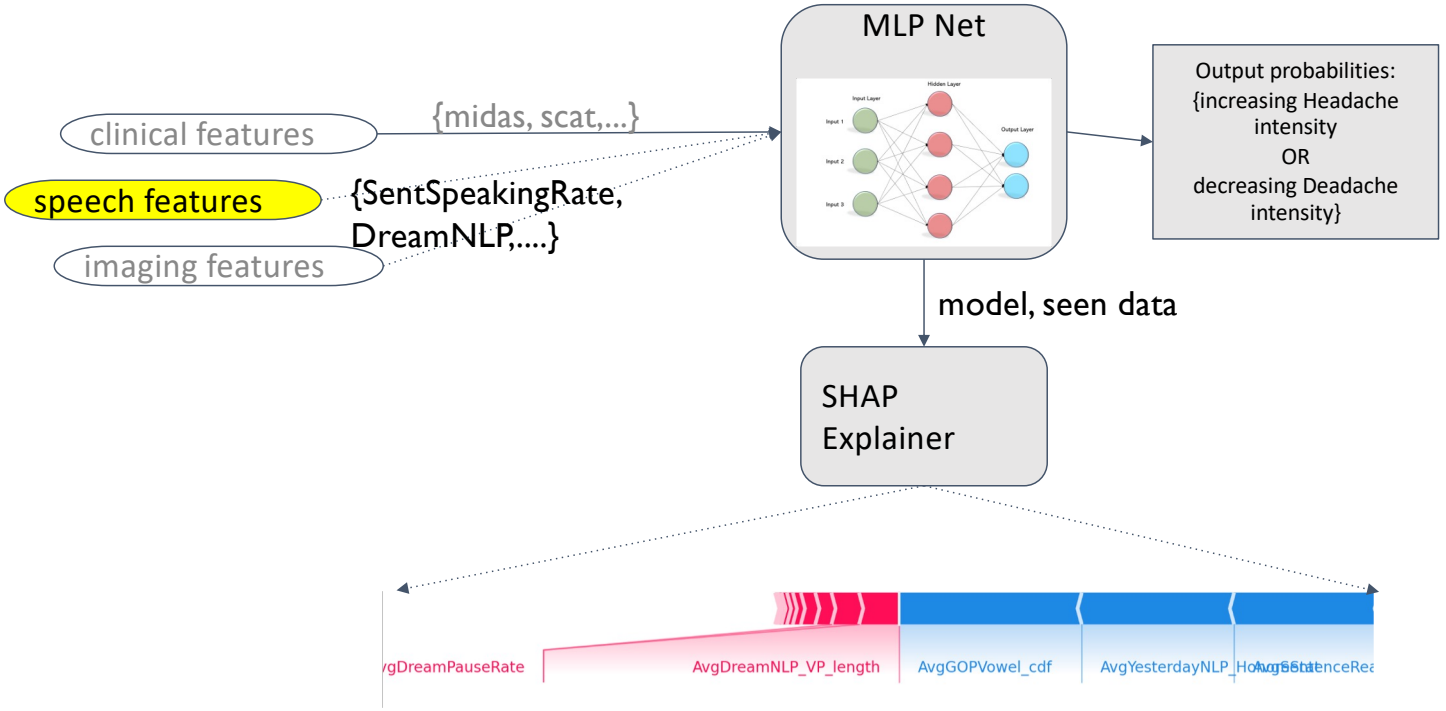
  - meaning → not all PTH patients have same reasons.

3. Speech data

Dataset used

- Speech data only from the NIH study
  - Dataset curation: Jianwei Zhang (ASU), Prof. Visar Berisha (ASU)
- Total sample size 22 patients
  - All 22 are PTH
  - 17 features
  - Acoustics:
    - GOPVowel, GOPConsonant (and their normalized values);
    - YesterdayPauseRate and DreamPauseRate;
    - SentenceReadingSpeakRate
  - Yesterday and Dream NLP:
    - BrunetIndex, HonoreStat, NP_rate, TTR, VBI/W, VP_length.

Overall Architecture (with speech data)



Observations

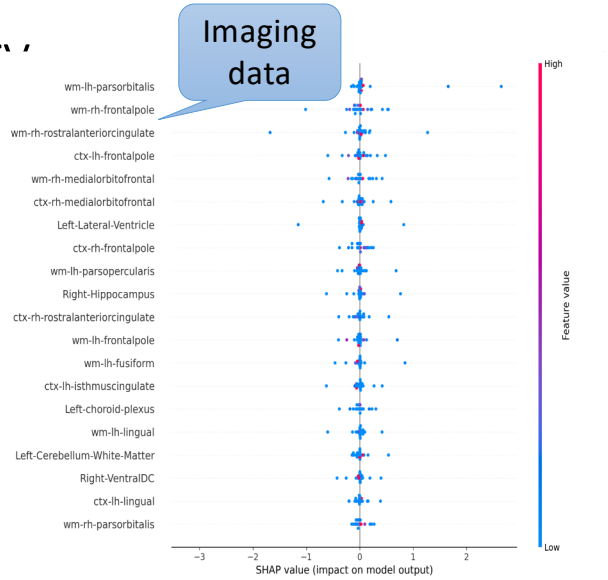
Overall (trained and tested on only PTH)		Ones with **increasing** headache intensity	
AvgSentenceReadingSpeakRate_cdf	9 out of 22	AvgSentenceReadingSpeakRate_cdf	4 out of 7
AvgDreamNLP_VP_length	7 out of 22	AvgYesterdayNLP_VP_length	2 out of 7
AvgYesterdayNLP_VP_length	6 out of 22	AvgDreamNLP_VP_length	2 out of 7
AvgYesterdayNLP_HonoreStat	6 out of 22	AvgDreamPauseRate	3 out of 7
AvgGOPVowel_cdf	6 out of 22	AvgYesterdayNLP_HonoreStat	2 out of 7
AvgDreamPauseRate	5 out of 22		
		Ones with **decreasing** headache intensity	
		AvgSentenceReadingSpeakRate_cdf	5 out of 15
		AvgDreamNLP_VP_length	5 out of 15
		AvgYesterdayNLP_VP_length	4 out of 15
		AvgYesterdayNLP_HonoreStat	4 out of 15
		AvgGOPVowel_cdf	5 out of 15
		AvgDreamPauseRate	2 out of 15

SHAP results for Biomarker Discover

Clinical data

IIHPTH001	scat_dont_feel	bpi_pain_severity_2	age	bpi_pain_severity_1	scat_num_symp
IIHPTH002	hh_ha_days_30	midas_3	scat_dont_feel	midas_ha_freq	brief_pain_inventory_7
IIHPTH003	scat_headache	age	bpi_pain_severity_4	scat_pressure	scat_neck_pain
IIHPTH004	oiss_concentrating	oiss_weakness	bpi_pain_severity_2	oiss_vision	gad_2_2
IIHPTH005	hh_ha_days_30	midas_3	bpi_pain_severity_1	phq_2_1	domain_pgic___5
IIHPTH006	trails_b_zscore	domain_pgic___5	oiss_concentrating	hyperacusis_score	scat_hamen
IIHPTH007	hh_ha_days_30	midas_3	scat_headache	midas_ha_freq	asc_score_ic
IIHPTH008	scat_headache	midas_3	domain_pgic___5	bdi_score	scat_pressure
IIHPTH009	midas_3	hh_ha_days_30	scat_headache	scat_dont_feel	brief_pain_inventory_7
IIHPTH010	ha_untreated_baseline	hh_congestion	domain_pgic___1	sleep_scale_11	hh_conjunct
IIHPTH011	ha_untreated_baseline	scat_dont_feel	asc_score_ic	scat_headache	bpi_pain_severity_4
	hh_ha_days_30	midas_ha_freq	scat_dont_feel	asc_score_ic	scat_headache
	scat_headache	bdi_score	age	asc_score_ic	domain_pgic___5
IIHPTH014	bdi_score	domain_pgic___4	insomnia_total	oiss_ohdas_points	compass_total_score
IIHPTH015	domain_pgic___6	bdi_score	scat_neck_pain	scat_haphys	oiss_concentrating
IIHPTH016	bdi_score	age	insomnia_total	domain_pgic___7	hh_congestion
IIHPTH017	midas_ha_freq	midas_3	scat_headache	hh_ha_days_30	scat_dont_feel
IIHPTH018	midas_ha_freq	hh_ha_days_30	scat_headache	scat_dont_feel	nsi_headaches
IIHPTH019	domain_pgic___5	sleep_scale_10	bdi_score	scat_remembering	oiss_ohdas_points
IIHPTH020	phq_2_1	domain_pgic___6	age	oiss_ohdas_points	gad_2_2
IIHPTH021	midas_3	hh_ha_days_30	midas_ha_freq	ha_untreated_baseline	asc_score_ic
IIHPTH022	midas_ha_freq	domain_pgic___6	hyperacusis_score	hh_ha_days_30	brief_pain_inventory_2
IIHPTH023	midas_3	asc_score_ic	scat_dont_feel	hh_ha_days_30	midas_ha_freq
IIHPTH024	midas_3	domain_pgic___4	ha_untreated_baseline	hh_ha_days_30	bdi_score
IIHVAPTH01	hh_ha_days_30	midas_ha_freq	scat_headache	asc_score_ic	hh_days_abort
IIHVAPTH02	scat_headache	midas_ha_freq	hh_ha_days_30	asc_score_ic	oiss_concentrating

Imaging data



Speech data

Overall (trained and tested on only PTH)		Ones with **increasing** headache intensity	
AvgSentenceReadingSpeakRate_cdf	9 out of 22	AvgSentenceReadingSpeakRate_cdf	4 out of 7
AvgDreamNLP_VP_length	7 out of 22	AvgYesterdayNLP_VP_length	2 out of 7
AvgYesterdayNLP_VP_length	6 out of 22	AvgDreamNLP_VP_length	2 out of 7
AvgYesterdayNLP_HonoreStat	6 out of 22	AvgDreamPauseRate	3 out of 7
AvgGOPVowel_cdf	6 out of 22	AvgYesterdayNLP_HonoreStat	2 out of 7
AvgDreamPauseRate	5 out of 22		
		Ones with **decreasing** headache intensity	
		AvgSentenceReadingSpeakRate_cdf	5 out of 15
		AvgDreamNLP_VP_length	5 out of 15
		AvgYesterdayNLP_VP_length	4 out of 15
		AvgYesterdayNLP_HonoreStat	4 out of 15
		AvgGOPVowel_cdf	5 out of 15
		AvgDreamPauseRate	2 out of 15

Outline of the talk

- Need for Explainability in Deep Learning
  - Landscape of methods
- Shapley values for explanation
- Research Problem
  - Results from SHAP
- **Next Steps**

Limitations

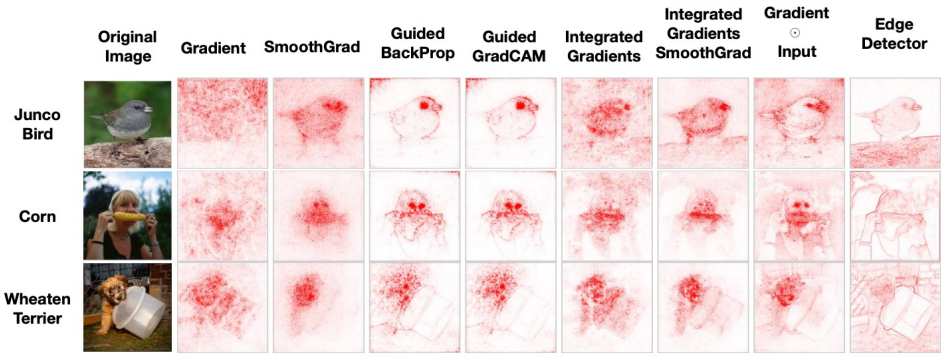
Truthfulness of explanations

Reliability of methods

Limitations

Truthfulness of explanations

Reliability of methods



Edge detectors are much similar to outputs of saliency maps on most methods

Limitations

Truthfulness of explanations

Reliability of methods

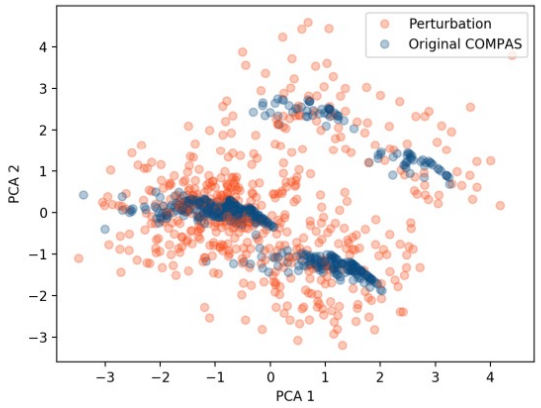


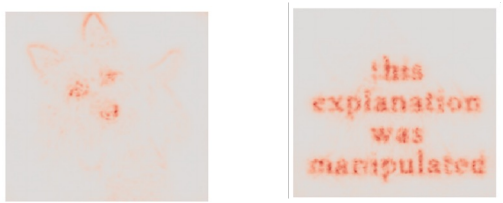
Figure 1: PCA applied to the COMPAS dataset (blue) as well as its LIME style perturbations (red). Even in this lov dimensional space, we can see that data points generate via perturbations are distributed very differently from ir stances in the COMPAS data. In this paper, we exploit thi difference to craft adversarial classifiers.

Perturbations generated in methods like LIME can be out of discribution of actual model's data cohort

And hence explanations generated from these methods can often fall into false conclusions

Hence, they are prone to "fooling" attacks

Post hoc explanations can be manipulated via adversarial attacks



What's next?

- Domain of Interpretable and Explainable AI is fairly new
- There's no one silver bullet
    - Definitions to explanations vary from application-to-application and user
- Research in post-hoc explainability is going to be more prevalent
- Focus on robust explainable systems
    - That cannot be hacked or fooled
- Metric to potentially evaluate performance of these methods
    - Or atleast some part of it

Thank You!

Questions?

Email: jgshah1@asu.edu
Homepage: https://jaygshah.github.io/