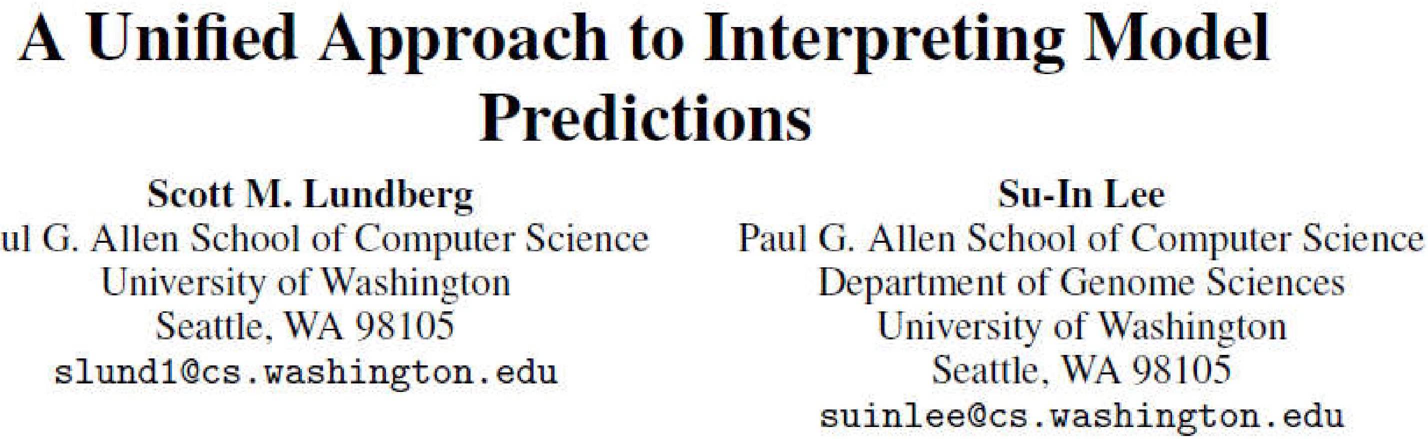
A Unified Approach to Interpreting Model Predictions: SHAP

CEB Journal club

Chaiyawat Suppasilp

19 Aug 2022

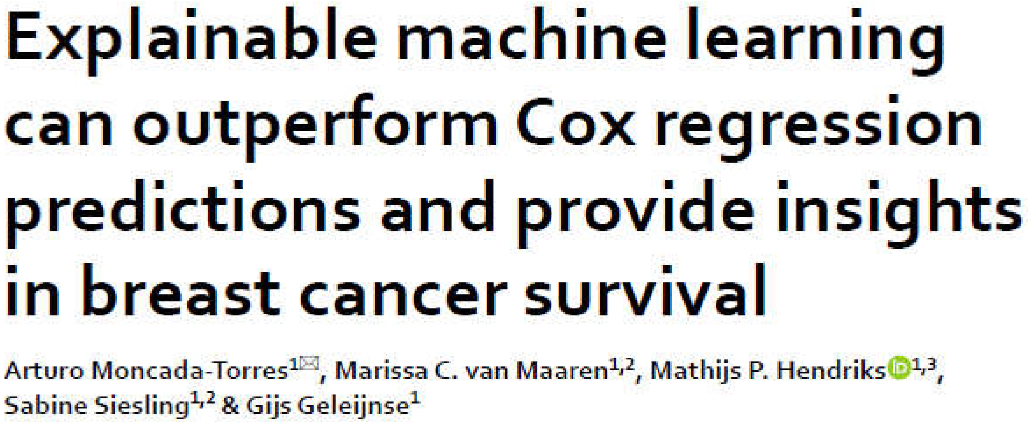
Journals



A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu



Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival

Arturo Moncada-Torres[1]✉, Marissa C. van Maaren[1,2], Mathijs P. Hendriks[1,3],
Sabine Siesling[1,2] & Gijs Geleijnse[1]

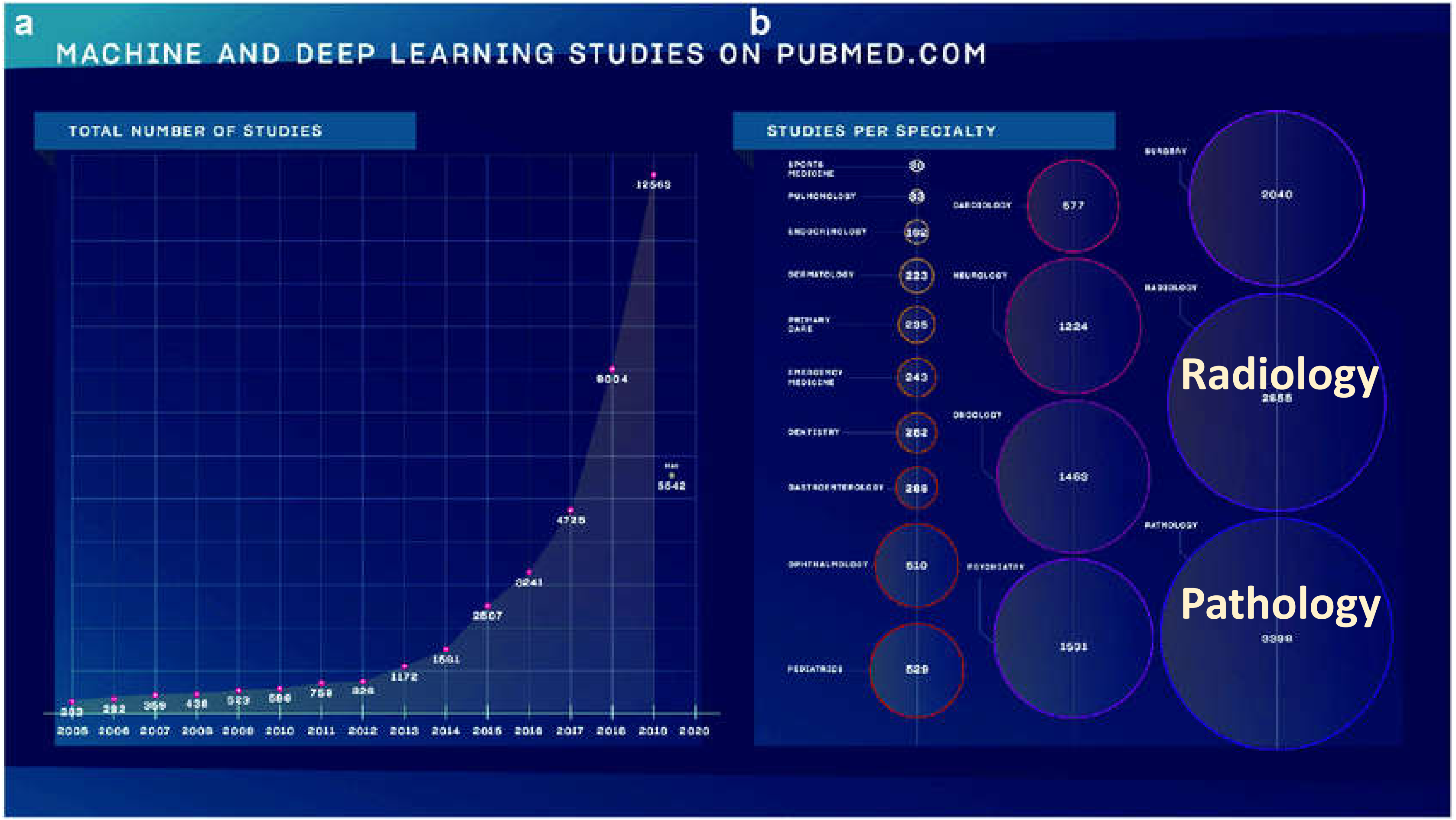
Acknowledgement

- Aj. Ratchainant Thammasudjarit
- Aditya Bhattacharya, Understand the Workings of SHAP and Shapley Values Used in Explainable AI
- Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable

Outline

- Introduction
- Shapley values
- SHAP values
- An example of using SHAP to provide insights in breast cancer survival
- Conclusion

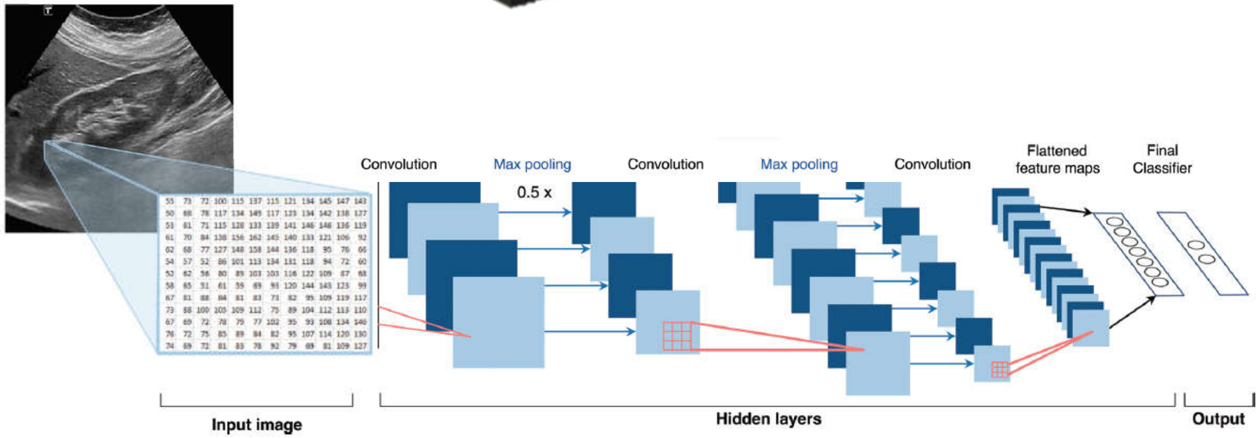
Rise of AI


a
b
MACHINE AND DEEP LEARNING STUDIES ON PUBMED.COM
TOTAL NUMBER OF STUDIES
203
282
359
438
523
596
758
828
1172
1581
2507
3241
4725
8004
12563
MAY
5542
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
STUDIES PER SPECIALTY
SPORTS MEDICINE
80
PULMONOLOGY
83
ENDOCRINOLOGY
192
DERMATOLOGY
223
PRIMARY CARE
235
EMERGENCY MEDICINE
243
DENTISTRY
262
GASTROENTEROLOGY
288
OPHTHALMOLOGY
510
PEDIATRICS
529
CARDIOLOGY
577
NEUROLOGY
1224
ONCOLOGY
1463
PSYCHIATRY
1531
SURGERY
2040
RADIOLOGY
Radiology
2655
PATHOLOGY
Pathology
3399

Black box

Input

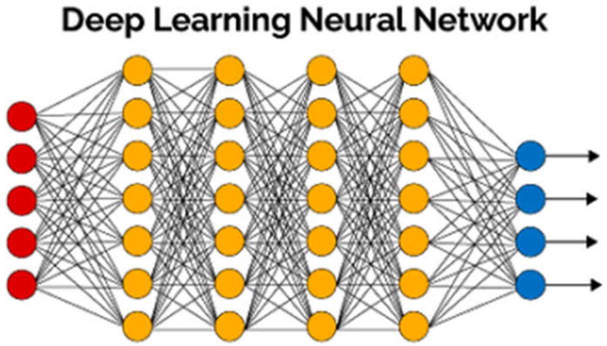
Output

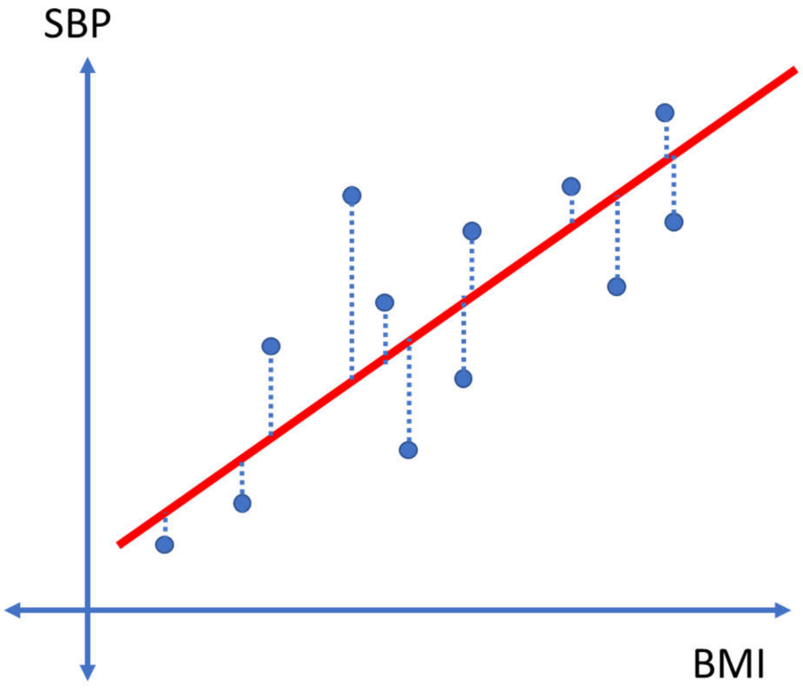


55	73	72	100	115	137	115	121	134	145	147	143
50	68	78	117	134	149	117	123	134	142	138	127
53	61	71	115	128	133	139	141	146	146	136	119
61	70	84	138	156	162	145	140	133	121	106	92
62	68	77	127	148	158	144	136	118	95	76	66
54	57	52	86	101	113	134	131	118	94	72	60
52	62	58	80	89	103	103	116	122	109	87	68
58	65	51	61	59	69	93	120	144	143	123	99
67	81	88	84	81	83	73	82	95	109	119	117
73	88	100	103	109	112	75	89	104	112	113	110
67	69	72	78	79	77	102	95	93	108	134	146
76	72	75	85	89	84	82	95	107	114	120	130
74	69	72	81	83	78	92	79	69	81	109	127

Transparency

- An understanding of how the algorithm works
- Least squares method for linear models
  → high transparency
- Deep learning approaches (gradient through a network with millions of weights)
  → less transparent



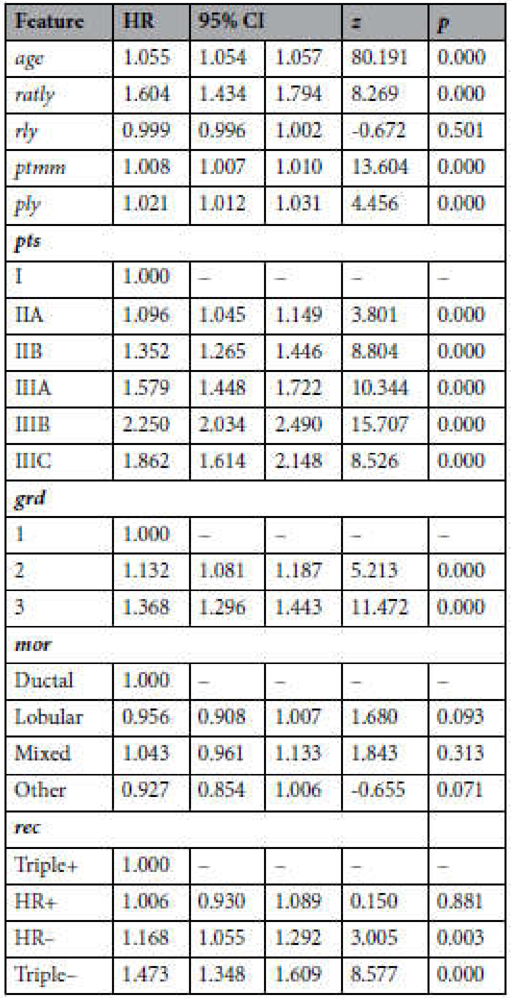


Least Sum of Squared Residual

$$\hat{y}_i = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + e_i$$

Interpretability

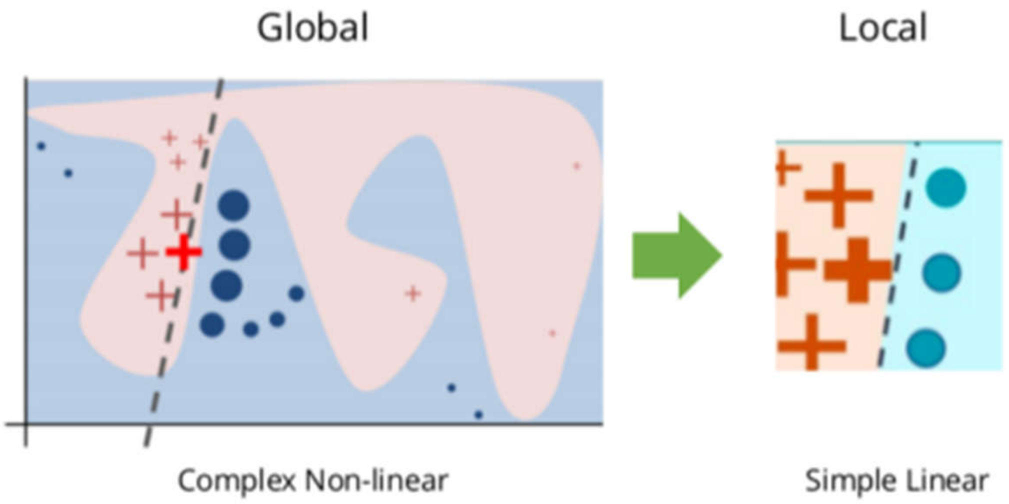
- Global Model Interpretability
  - How does the trained model make predictions?
  - Understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures
  - A Modular Level
    = How do parts of the model affect predictions?
    → single weight or a coefficient of regression model

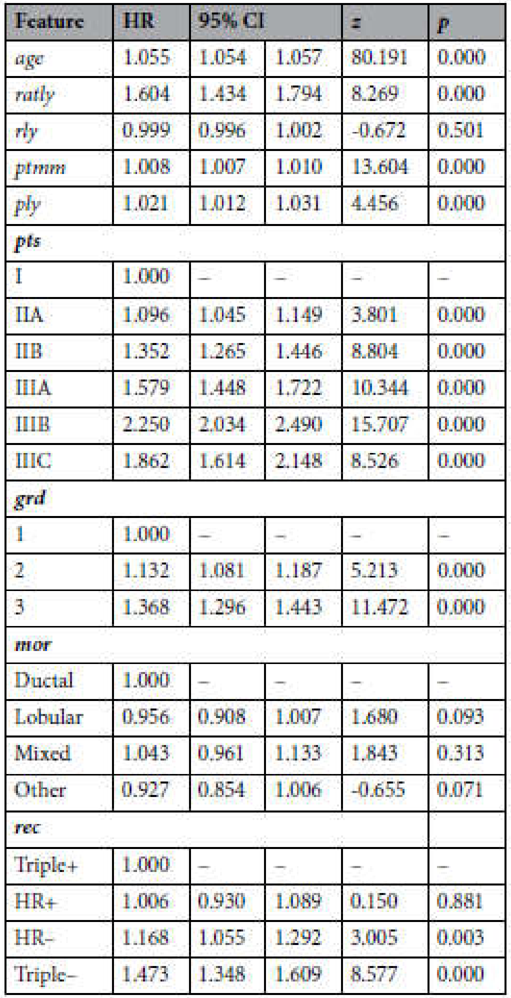


Feature	HR	95% CI		z	p
age	1.055	1.054	1.057	80.191	0.000
ratly	1.604	1.434	1.794	8.269	0.000
rly	0.999	0.996	1.002	-0.672	0.501
ptmm	1.008	1.007	1.010	13.604	0.000
ply	1.021	1.012	1.031	4.456	0.000
pts					
I	1.000	–	–	–	–
IIA	1.096	1.045	1.149	3.801	0.000
IIB	1.352	1.265	1.446	8.804	0.000
IIIA	1.579	1.448	1.722	10.344	0.000
IIIB	2.250	2.034	2.490	15.707	0.000
IIIC	1.862	1.614	2.148	8.526	0.000
grd					
1	1.000	–	–	–	–
2	1.132	1.081	1.187	5.213	0.000
3	1.368	1.296	1.443	11.472	0.000
mor					
Ductal	1.000	–	–	–	–
Lobular	0.956	0.908	1.007	1.680	0.093
Mixed	1.043	0.961	1.133	1.843	0.313
Other	0.927	0.854	1.006	-0.655	0.071
rec					
Triple+	1.000	–	–	–	–
HR+	1.006	0.930	1.089	0.150	0.881
HR–	1.168	1.055	1.292	3.005	0.003
Triple–	1.473	1.348	1.609	8.577	0.000

Interpretability

- Local Interpretability for a Single Prediction
  - Why did the model make a certain prediction for an instance?



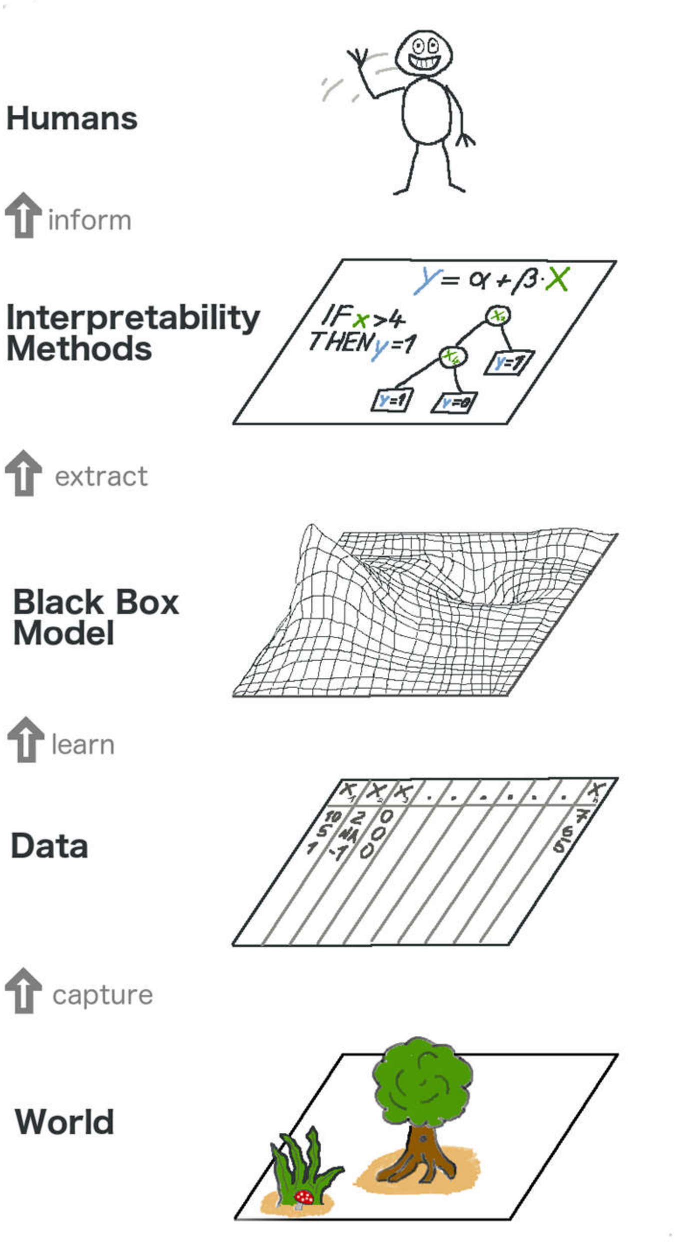


Feature	HR	95% CI		z	p
age	1.055	1.054	1.057	80.191	0.000
ratly	1.604	1.434	1.794	8.269	0.000
rly	0.999	0.996	1.002	-0.672	0.501
ptmm	1.008	1.007	1.010	13.604	0.000
ply	1.021	1.012	1.031	4.456	0.000
pts					
I	1.000	–	–	–	–
IIA	1.096	1.045	1.149	3.801	0.000
IIB	1.352	1.265	1.446	8.804	0.000
IIIA	1.579	1.448	1.722	10.344	0.000
IIIB	2.250	2.034	2.490	15.707	0.000
IIIC	1.862	1.614	2.148	8.526	0.000
grd					
1	1.000	–	–	–	–
2	1.132	1.081	1.187	5.213	0.000
3	1.368	1.296	1.443	11.472	0.000
mor					
Ductal	1.000	–	–	–	–
Lobular	0.956	0.908	1.007	1.680	0.093
Mixed	1.043	0.961	1.133	1.843	0.313
Other	0.927	0.854	1.006	-0.655	0.071
rec					
Triple+	1.000	–	–	–	–
HR+	1.006	0.930	1.089	0.150	0.881
HR–	1.168	1.055	1.292	3.005	0.003
Triple–	1.473	1.348	1.609	8.577	0.000

ISSUE

The transparency and interpretability are the huge problem for applying AI in the real-world practice.

(but today, we focus on interpretability)



Humans

inform

Interpretability Methods

extract

Black Box Model

learn

Data

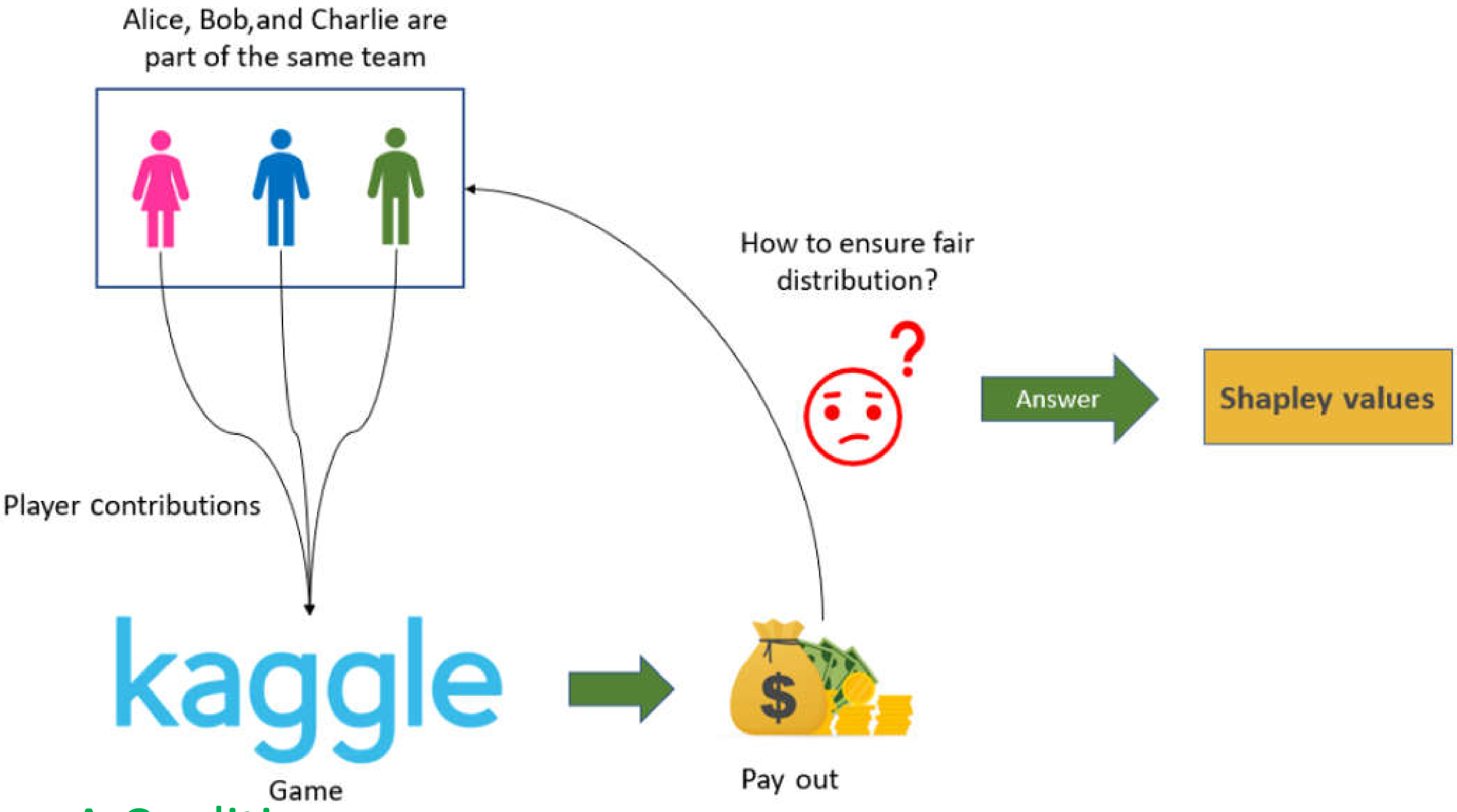
capture

World

How to solve the problems?

- Global Model-Agnostic Methods
  - The partial dependence plot is a feature effect method.
  - Permutation feature importance measures the importance of a feature as an increase in loss when the feature is permuted.
- Local Model-Agnostic Methods
  - Local surrogate models (LIME) explain a prediction by replacing the complex model with a locally interpretable surrogate model.
  - Shapley values are an attribution method that fairly assigns the prediction to individual features.
  - SHAP is another computation method for Shapley values, but also proposes global interpretation methods based on combinations of Shapley values across the data.

Shapley value was introduced in 1951 by Lloyd Shapley.



Alice, Bob,and Charlie are part of the same team

Player contributions

Game

Pay out

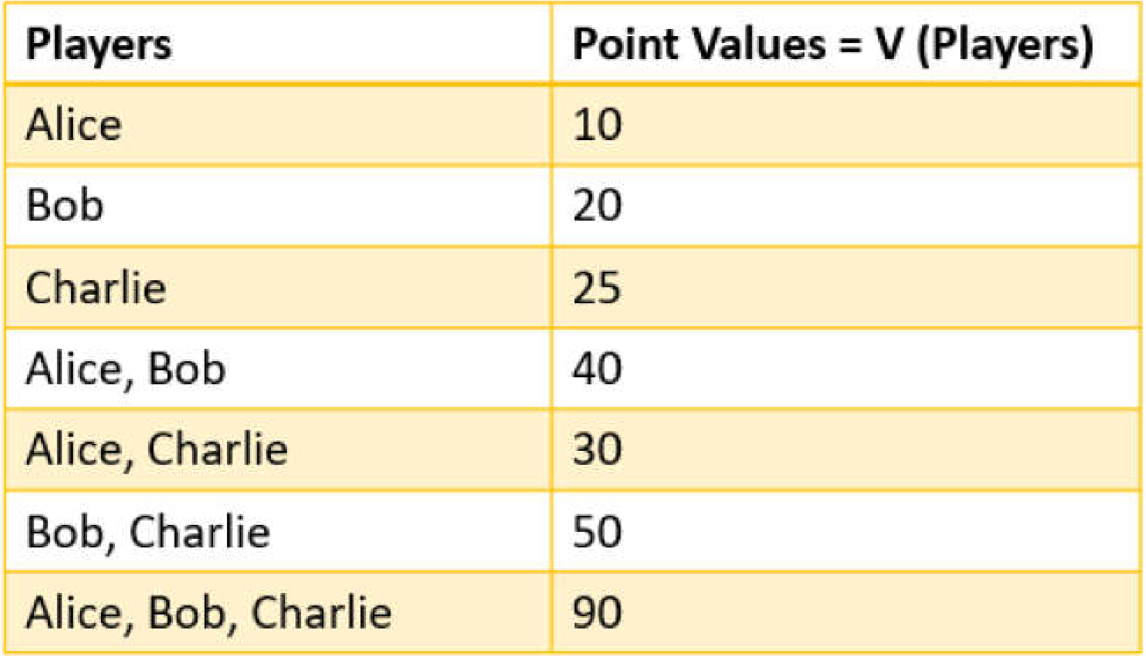
How to ensure fair distribution?

Answer

Shapley values

A Coalition game

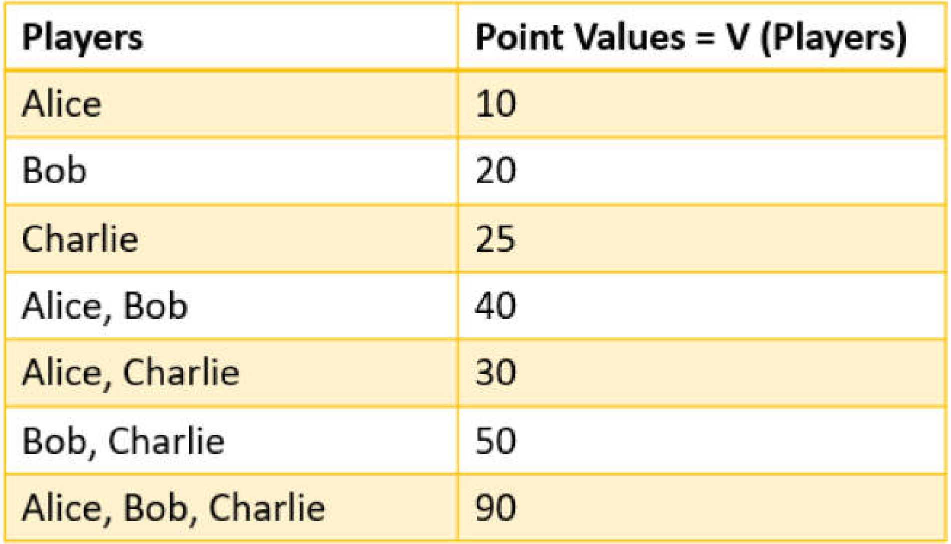
Shapley value



Players	Point Values = V (Players)
Alice	10
Bob	20
Charlie	25
Alice, Bob	40
Alice, Charlie	30
Bob, Charlie	50
Alice, Bob, Charlie	90

This table illustrates the point values for each condition for calculating the average marginal contribution of each player

Shapley value



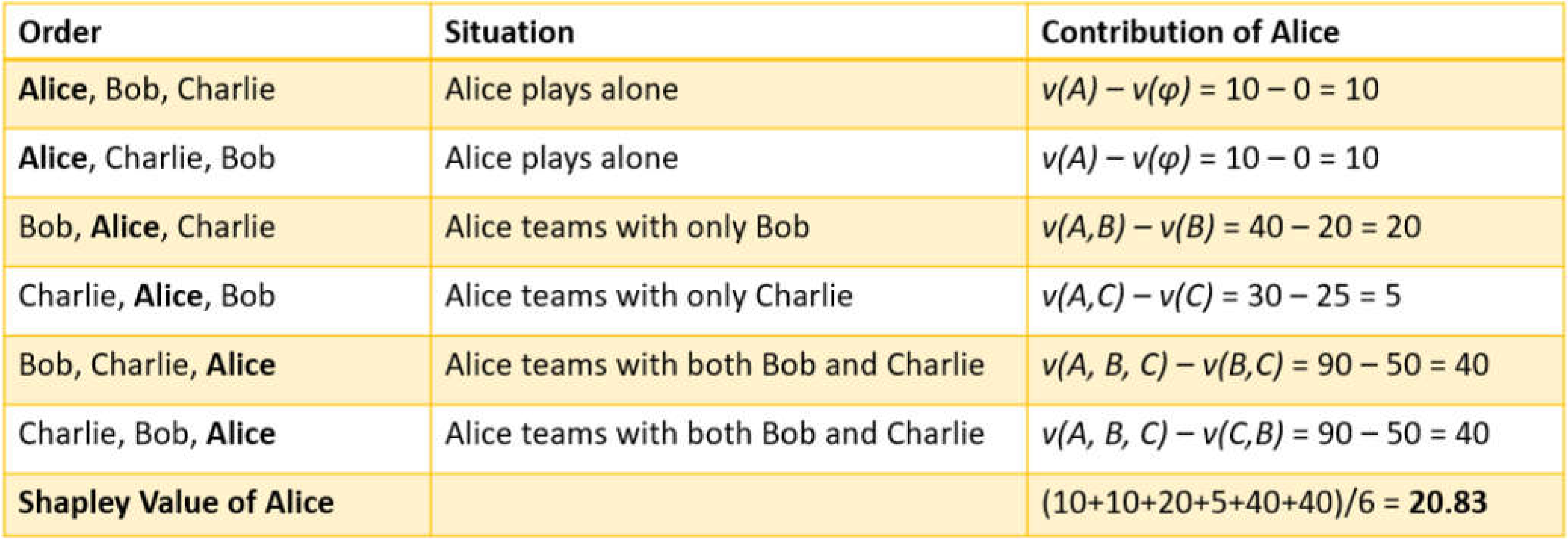
Players	Point Values = V (Players)
Alice	10
Bob	20
Charlie	25
Alice, Bob	40
Alice, Charlie	30
Bob, Charlie	50
Alice, Bob, Charlie	90

$$\varphi(i) = \sum_{S\subseteq N/i} \frac{|S|!\,(|N|-|S|-1)!}{|N|!}\left(v(S\ \cup\ \{i\}) - v(S)\right)$$

N	Number of players
S	Coalition subset of players
$v(S)$	Total value of S players
i	Individual player
$\varphi(i)$	Marginal contribution of player i

Shapley value

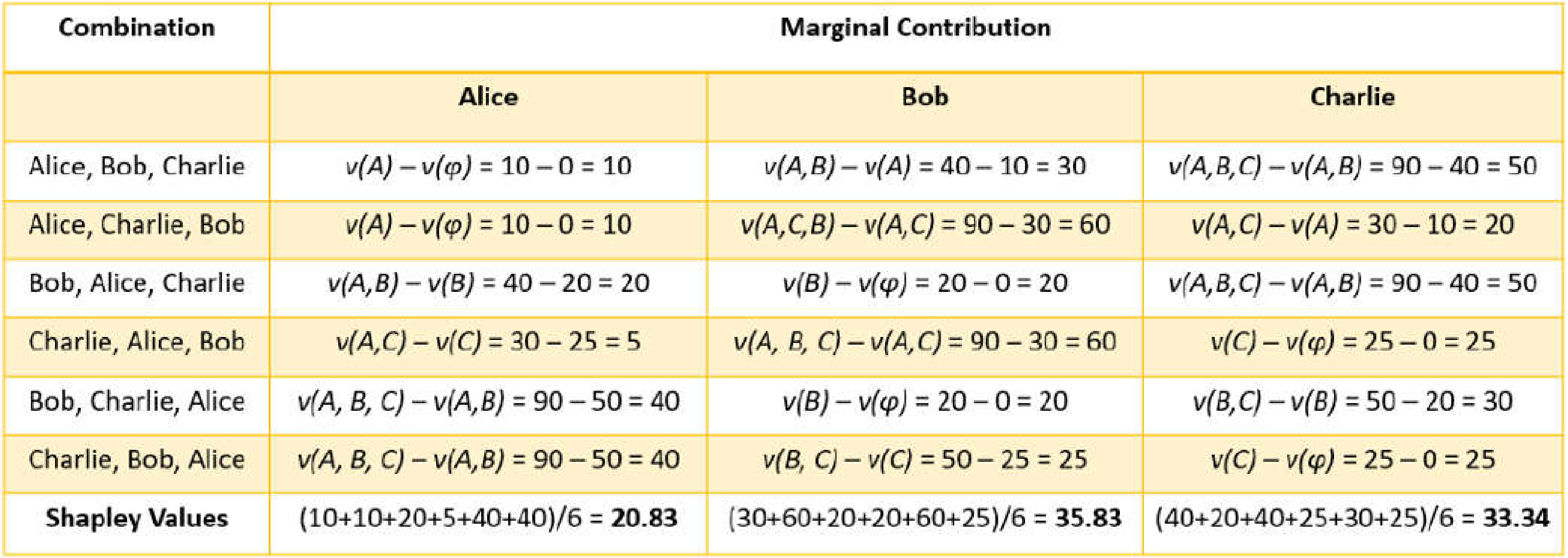
$$\varphi(i) = \sum_{S\subseteq N/i} \frac{|S|!\,(|N|-|S|-1)!}{|N|!}\left(v(S \cup \{i\}) - v(S)\right)$$



Order	Situation	Contribution of Alice
Alice, Bob, Charlie	Alice plays alone	*v(A) – v(φ)* = 10 – 0 = 10
Alice, Charlie, Bob	Alice plays alone	*v(A) – v(φ)* = 10 – 0 = 10
Bob, **Alice**, Charlie	Alice teams with only Bob	*v(A,B) – v(B)* = 40 – 20 = 20
Charlie, **Alice**, Bob	Alice teams with only Charlie	*v(A,C) – v(C)* = 30 – 25 = 5
Bob, Charlie, **Alice**	Alice teams with both Bob and Charlie	*v(A, B, C) – v(B,C)* = 90 – 50 = 40
Charlie, Bob, **Alice**	Alice teams with both Bob and Charlie	*v(A, B, C) – v(C,B)* = 90 – 50 = 40
Shapley Value of Alice		(10+10+20+5+40+40)/6 = **20.83**

The Shapley value of Alice is her marginal contribution considering all possible scenarios

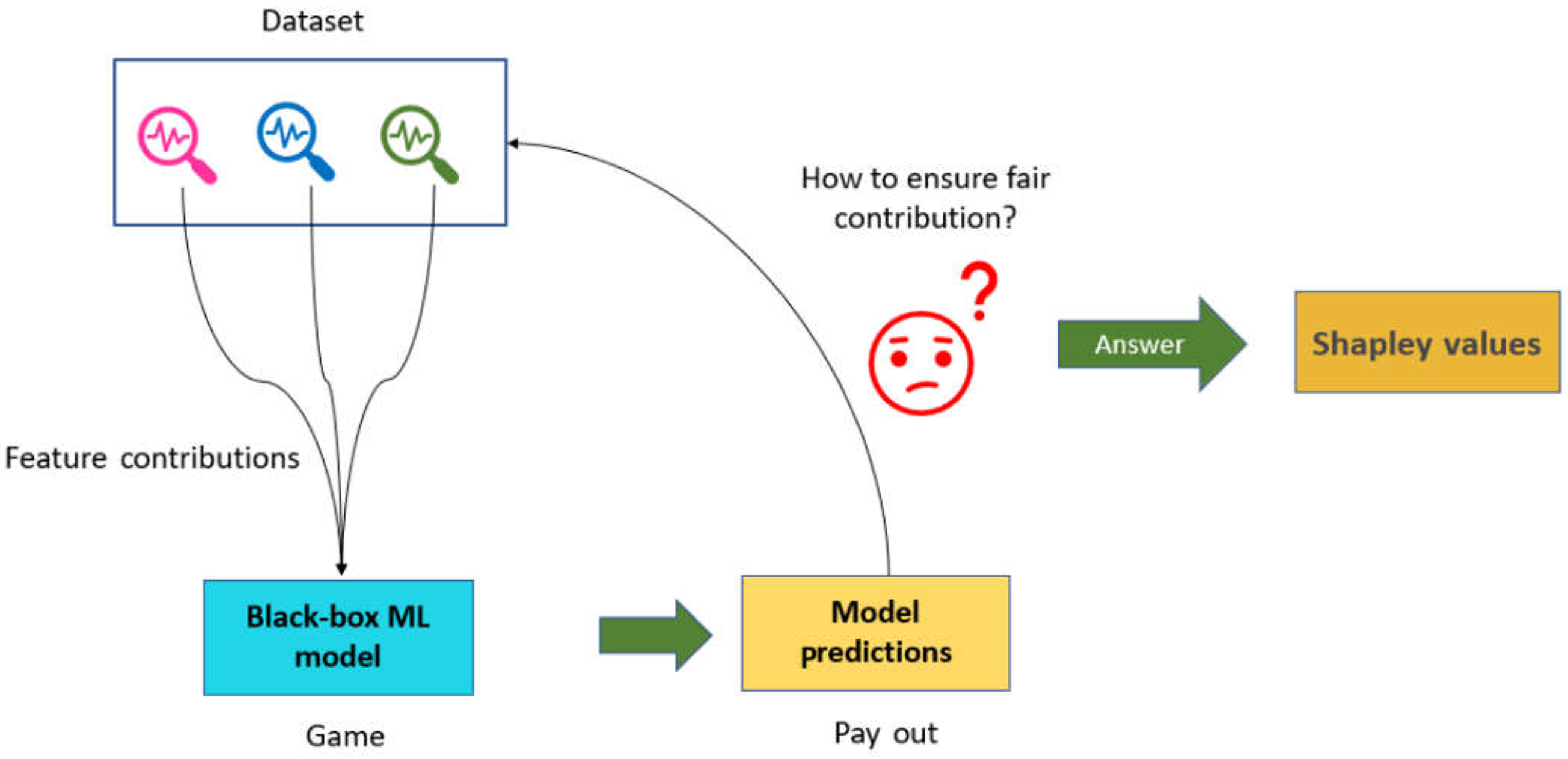
Shapley value



Combination	Marginal Contribution		
	Alice	**Bob**	**Charlie**
Alice, Bob, Charlie	$v(A) - v(\varphi) = 10 - 0 = 10$	$v(A,B) - v(A) = 40 - 10 = 30$	$v(A,B,C) - v(A,B) = 90 - 40 = 50$
Alice, Charlie, Bob	$v(A) - v(\varphi) = 10 - 0 = 10$	$v(A,C,B) - v(A,C) = 90 - 30 = 60$	$v(A,C) - v(A) = 30 - 10 = 20$
Bob, Alice, Charlie	$v(A,B) - v(B) = 40 - 20 = 20$	$v(B) - v(\varphi) = 20 - 0 = 20$	$v(A,B,C) - v(A,B) = 90 - 40 = 50$
Charlie, Alice, Bob	$v(A,C) - v(C) = 30 - 25 = 5$	$v(A, B, C) - v(A,C) = 90 - 30 = 60$	$v(C) - v(\varphi) = 25 - 0 = 25$
Bob, Charlie, Alice	$v(A, B, C) - v(A,B) = 90 - 50 = 40$	$v(B) - v(\varphi) = 20 - 0 = 20$	$v(B,C) - v(B) = 50 - 20 = 30$
Charlie, Bob, Alice	$v(A, B, C) - v(A,B) = 90 - 50 = 40$	$v(B, C) - v(C) = 50 - 25 = 25$	$v(C) - v(\varphi) = 25 - 0 = 25$
Shapley Values	(10+10+20+5+40+40)/6 = **20.83**	(30+60+20+20+60+25)/6 = **35.83**	(40+20+40+25+30+25)/6 = **33.34**

Marginal contribution for Alice, Bob, and Charlie

Shapley value



Advantages of Shapley value

- Fair payout: Efficiency, Symmetry, Dummy and Additivity
  (Fair payout for all features, NOT like LIME)
- Contrastive explanations: Instead of comparing a prediction to the average prediction of the entire dataset, you could compare it to a subset or even to a single data point.
- Solid theory

Disadvantages of Shapley value

- Computing time
- Explanations created with the Shapley value method always use all the features.
  (SHAP can provide explanations with few features)
- No prediction model
  (ex. If age is increased by 1 year, the risk will be increased 5%)
- You need access to the data if you want to calculate the Shapley value for a new data instance.

SHAP value

- SHapley Additive exPlanations
- by Lundberg and Lee (2017)
- Shapley values are the only solution that satisfies properties of Efficiency, Symmetry, Dummy and Additivity.

- Proposed KernelSHAP, an alternative, kernel-based estimation approach for Shapley values inspired by local surrogate models.
- SHAP comes with many global interpretation methods based on aggregations of Shapley values
- SHAP has both global and local interpretability

Simple Properties Uniquely Determine Additive Feature Attributions

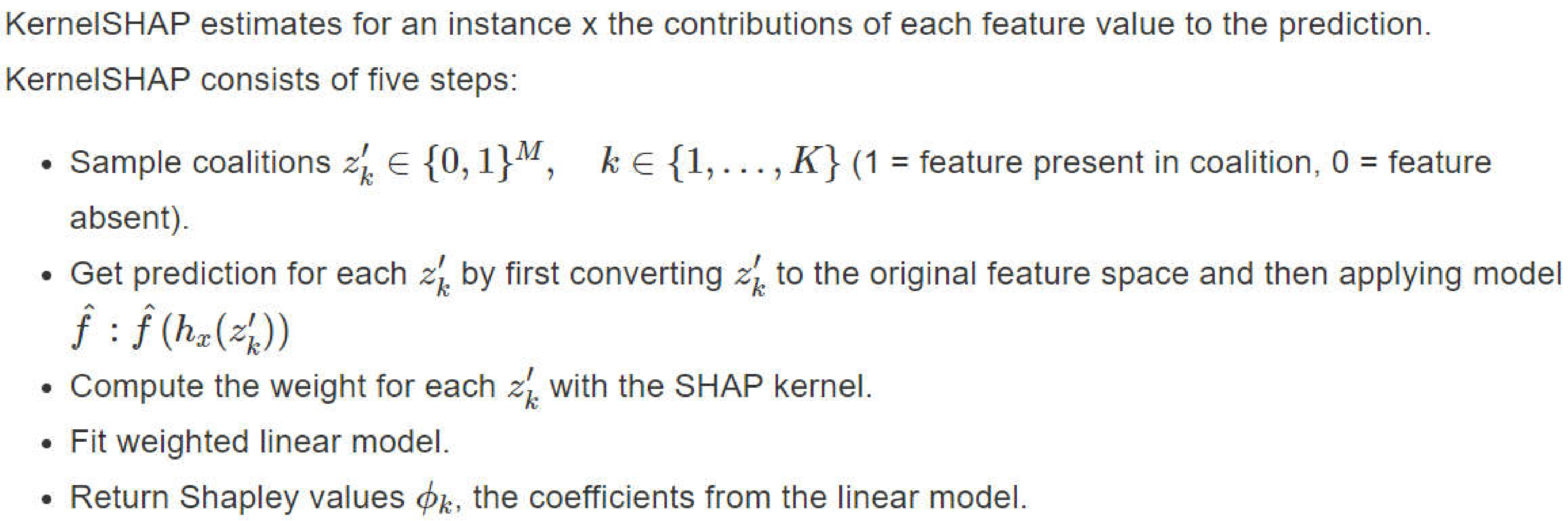
1. Local accuracy ≈ Shapley efficiency property
2. Missingness : If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact.
3. Consistency: if a model changes so that the marginal contribution of a feature value increases or stays the same (regardless of other features and not decreased), the Shapley value also increases or stays the same. F

SHAP value

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

g	the explanation model
$z' \in \{0,1\}^M$	the coalition vector or simplified features
M	the maximum coalition size
$\phi_j \in \mathbb{R}$	the feature attribution for a feature j, the Shapley values

Kernel SHAP



KernelSHAP estimates for an instance x the contributions of each feature value to the prediction. KernelSHAP consists of five steps:

- Sample coalitions $z'_k \in \{0,1\}^M, \quad k \in \{1,\ldots,K\}$ (1 = feature present in coalition, 0 = feature absent).
- Get prediction for each z'_k by first converting z'_k to the original feature space and then applying model $\hat{f} : \hat{f}(h_x(z'_k))$
- Compute the weight for each z'_k with the SHAP kernel.
- Fit weighted linear model.
- Return Shapley values ϕ_k, the coefficients from the linear model.

Kernel SHAP

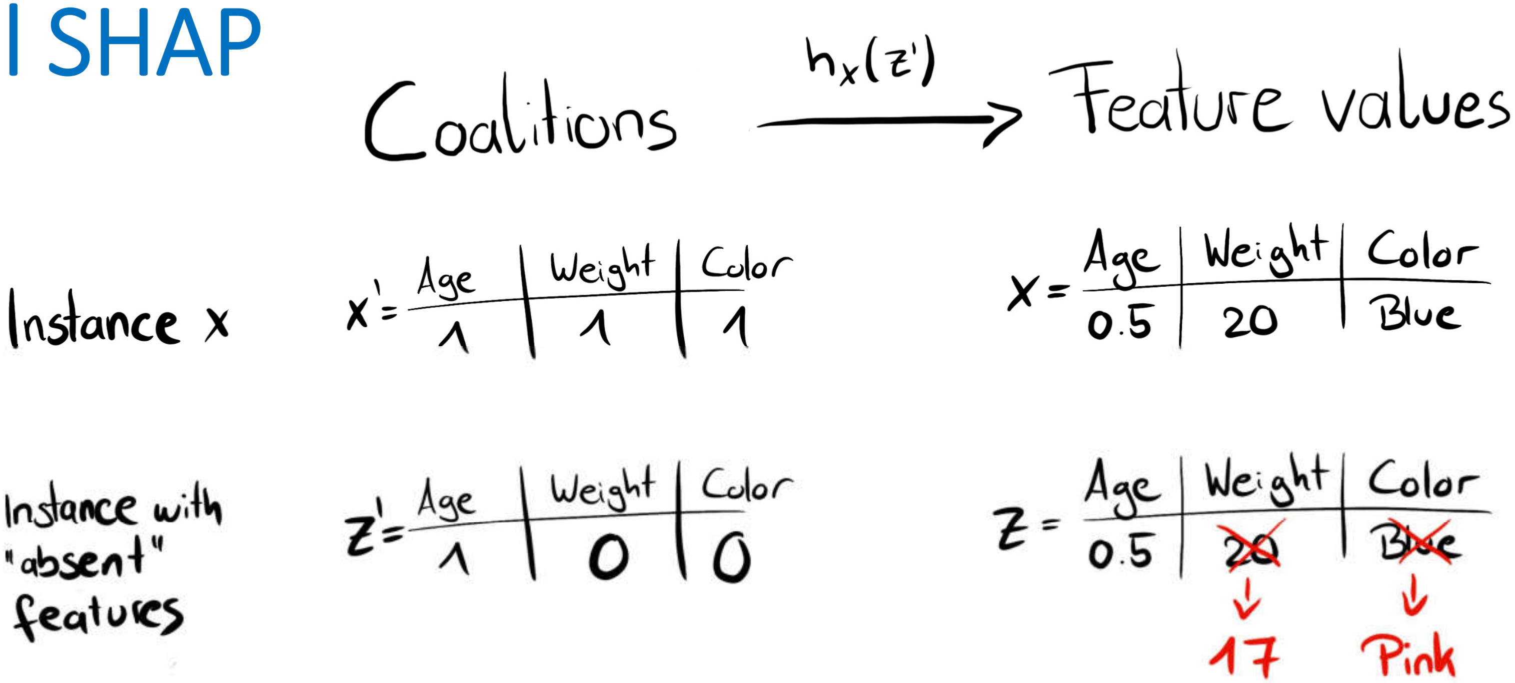
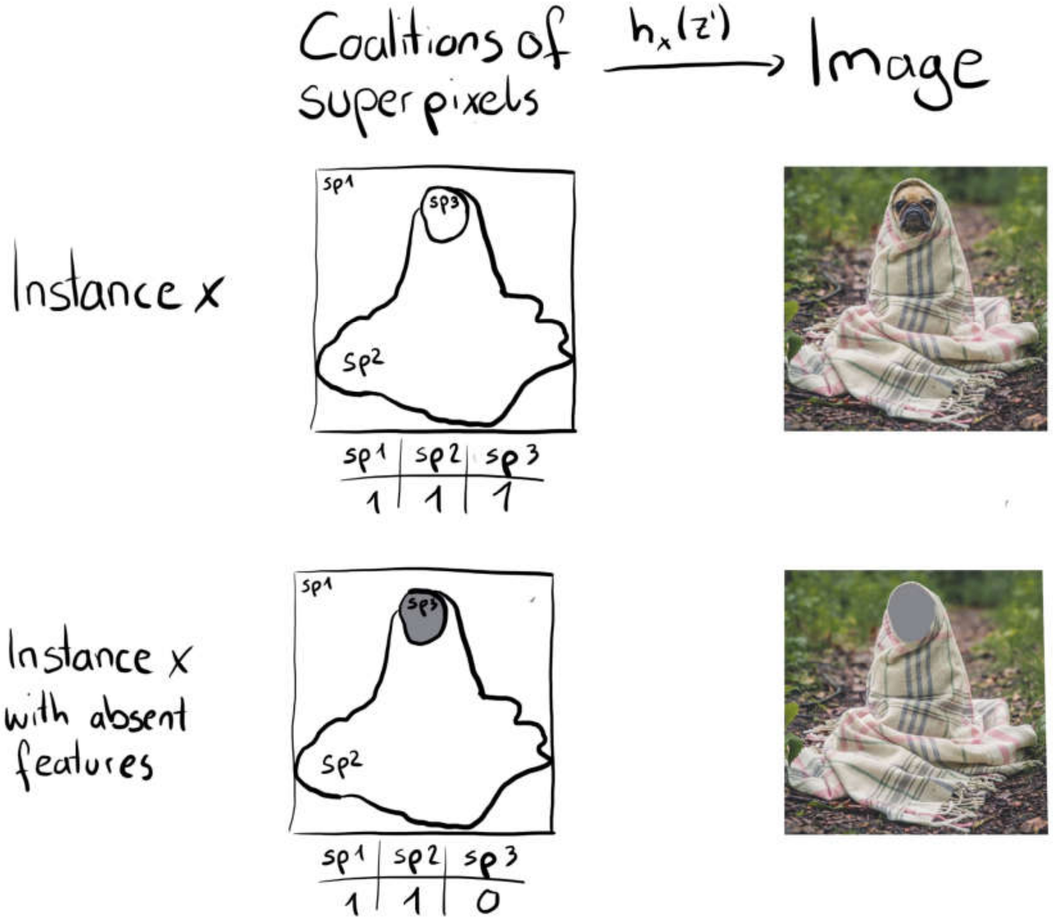


FIGURE 9.22: Function h_x maps a coalition to a valid instance. For present features (1), h_x maps to the feature values of x. For absent features (0), h_x maps to the values of a randomly sampled data instance.

h_x for tabular data treats X_C and X_S as independent and integrates over the marginal distribution:

$$\hat{f}\,(h_x(z')) = E_{X_C}[\hat{f}\,(x)]$$

Kernel SHAP



sp1	sp2	sp3
1	1	1

sp1	sp2	sp3
1	1	0

FIGURE 9.23: Function h_x maps coalitions of superpixels (sp) to images. Superpixels are groups of pixels. For present features (1), h_x returns the corresponding part of the original image. For absent features (0), h_x greys out the corresponding area. Assigning the average color of surrounding pixels or similar would also be an option.

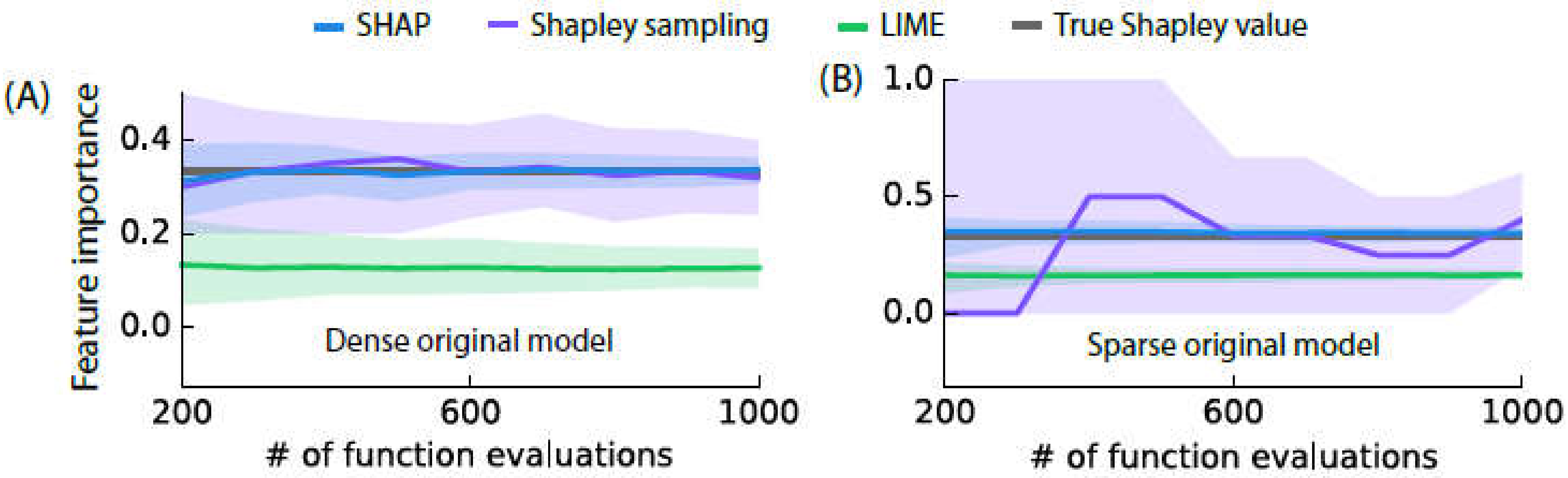
Kernel SHAP

- Lundberg et al. propose the SHAP kernel:

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

- M is the maximum coalition size and |z′ the number of present features in instance z'. Lundberg and Lee show that linear regression with this kernel weight yields Shapley values.

Computational Efficiency



Feature importance estimates

(A) A decision tree model using all 10 input features is explained for a single input.

(B) A decision tree using only 3 of 100 input features is explained for a single input.

Consistency with Human Intuition

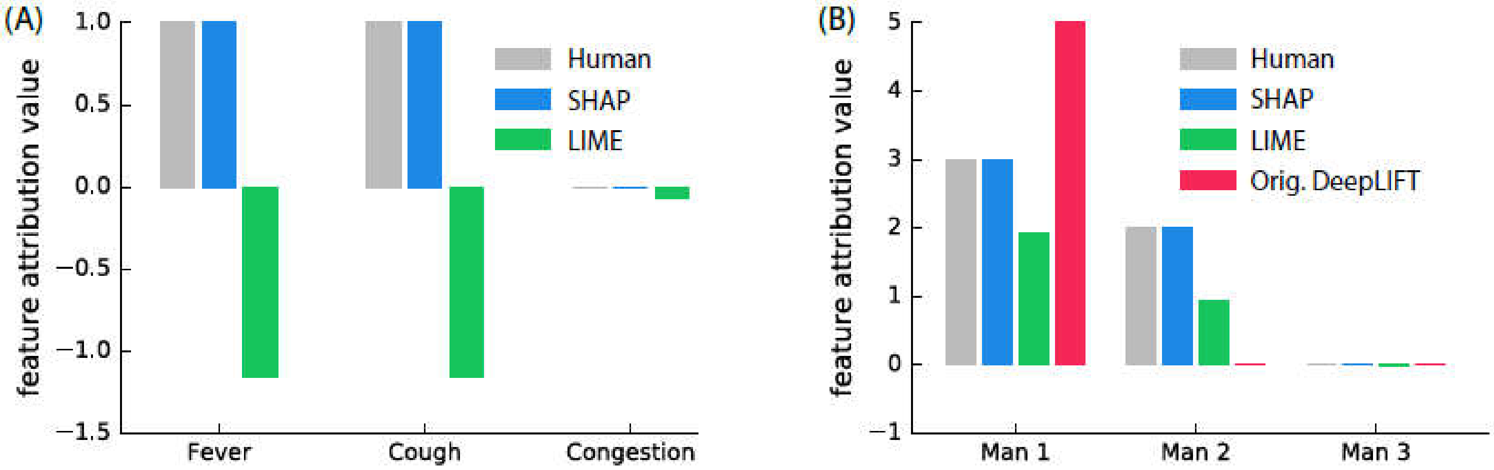


Figure 4: Human feature impact estimates are shown as the most common explanation given among 30 (A) and 52 (B) random individuals, respectively. (A) Feature attributions for a model output value (sickness score) of 2. The model output is 2 when fever and cough are both present, 5 when only one of fever or cough is present, and 0 otherwise. (B) Attributions of profit among three men, given according to the maximum number of questions any man got right. The first man got 5 questions right, the second 4 questions, and the third got none right, so the profit is $5.

Explaining Class Differences

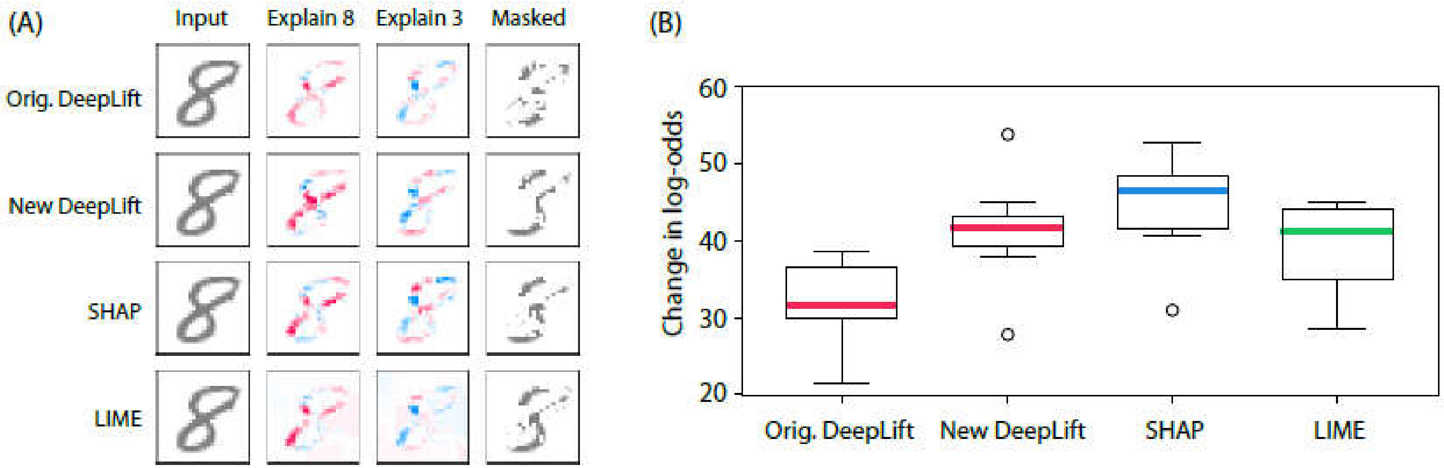
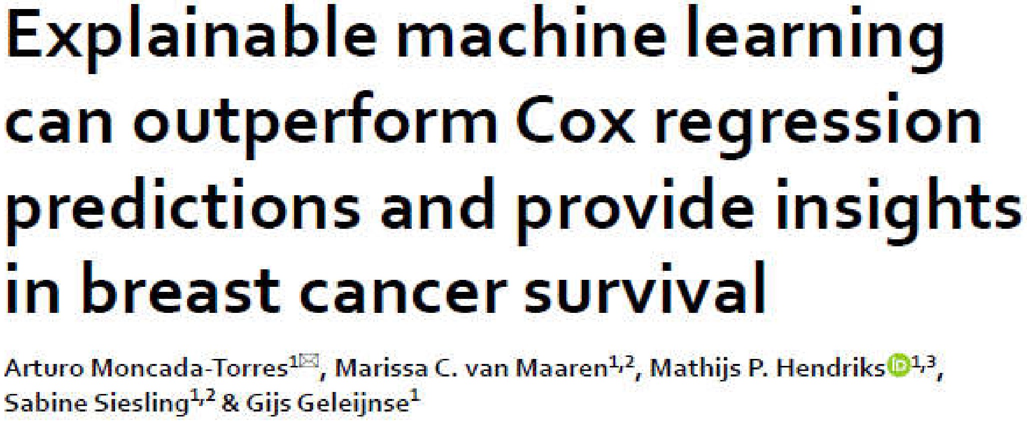


Figure 5: Explaining the output of a convolutional network trained on the MNIST digit dataset. Orig. DeepLIFT has no explicit Shapley approximations, while New DeepLIFT seeks to better approximate Shapley values. (A) Red areas increase the probability of that class, and blue areas decrease the probability. Masked removes pixels in order to go from 8 to 3. (B) The change in log odds when masking over 20 random images supports the use of better estimates of SHAP values.

Advantages of SHAP

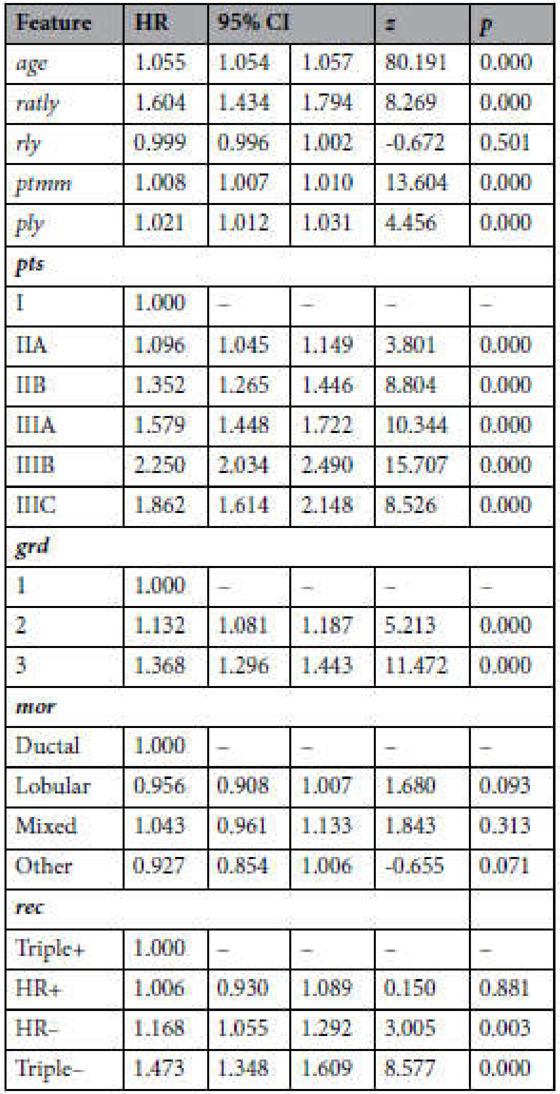
- Global interpretability: SHAP values not only show feature importance but also show whether the feature has a positive or negative impact on predictions.
- Local interpretability: We can calculate SHAP values for each individual prediction and know how the features contribute to that single prediction. Other techniques only show aggregated results over the whole dataset.
- SHAP values can be used to explain a large variety of models including linear models (e.g. linear regression), tree-based models (e.g. XGBoost) and neural networks, while other techniques can only be used to explain limited model types.

Example



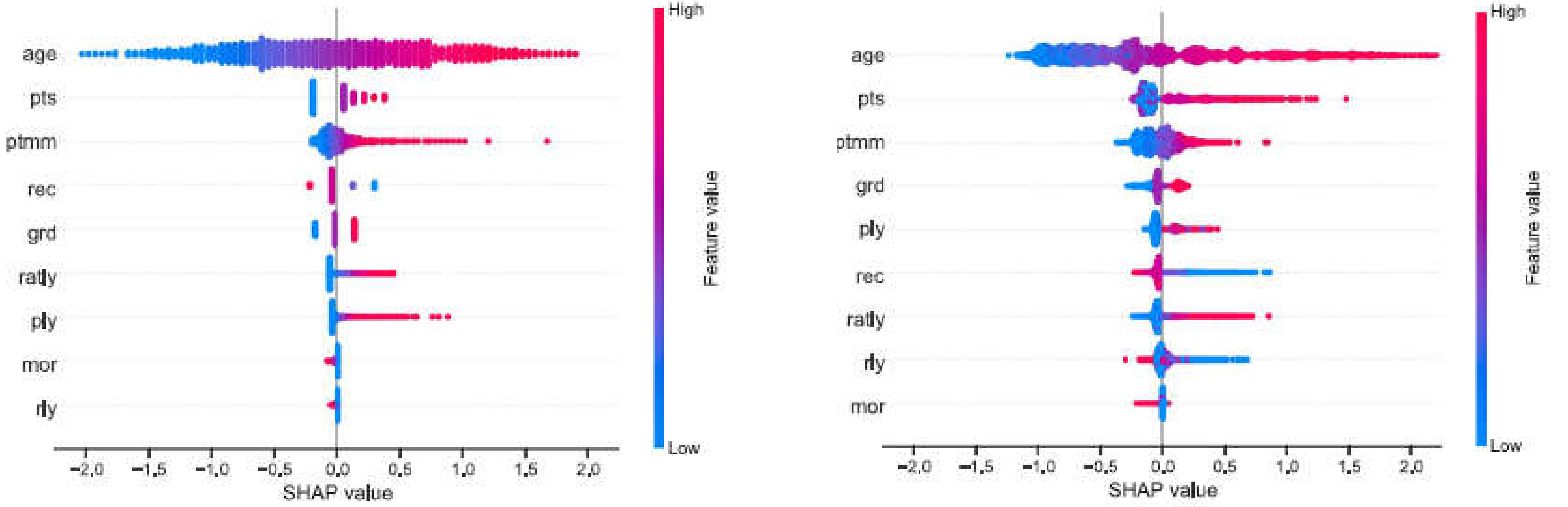
Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival

Arturo Moncada-Torres[1]✉, Marissa C. van Maaren[1,2], Mathijs P. Hendriks[1,3], Sabine Siesling[1,2] & Gijs Geleijnse[1]



Feature	HR	95% CI		z	p
age	1.055	1.054	1.057	80.191	0.000
ratly	1.604	1.434	1.794	8.269	0.000
rly	0.999	0.996	1.002	-0.672	0.501
ptmm	1.008	1.007	1.010	13.604	0.000
ply	1.021	1.012	1.031	4.456	0.000
pts					
I	1.000	–	–	–	–
IIA	1.096	1.045	1.149	3.801	0.000
IIB	1.352	1.265	1.446	8.804	0.000
IIIA	1.579	1.448	1.722	10.344	0.000
IIIB	2.250	2.034	2.490	15.707	0.000
IIIC	1.862	1.614	2.148	8.526	0.000
grd					
1	1.000	–	–	–	–
2	1.132	1.081	1.187	5.213	0.000
3	1.368	1.296	1.443	11.472	0.000
mor					
Ductal	1.000	–	–	–	–
Lobular	0.956	0.908	1.007	1.680	0.093
Mixed	1.043	0.961	1.133	1.843	0.313
Other	0.927	0.854	1.006	-0.655	0.071
rec					
Triple+	1.000	–	–	–	–
HR+	1.006	0.930	1.089	0.150	0.881
HR–	1.168	1.055	1.292	3.005	0.003
Triple–	1.473	1.348	1.609	8.577	0.000

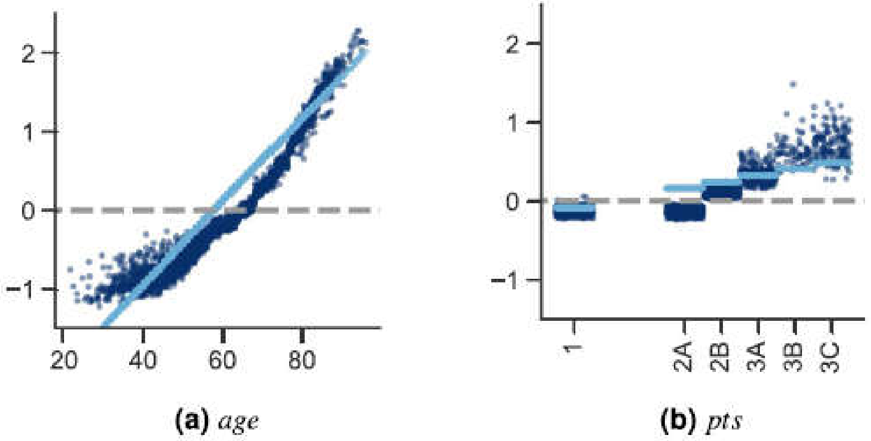
Example



(a) CPH

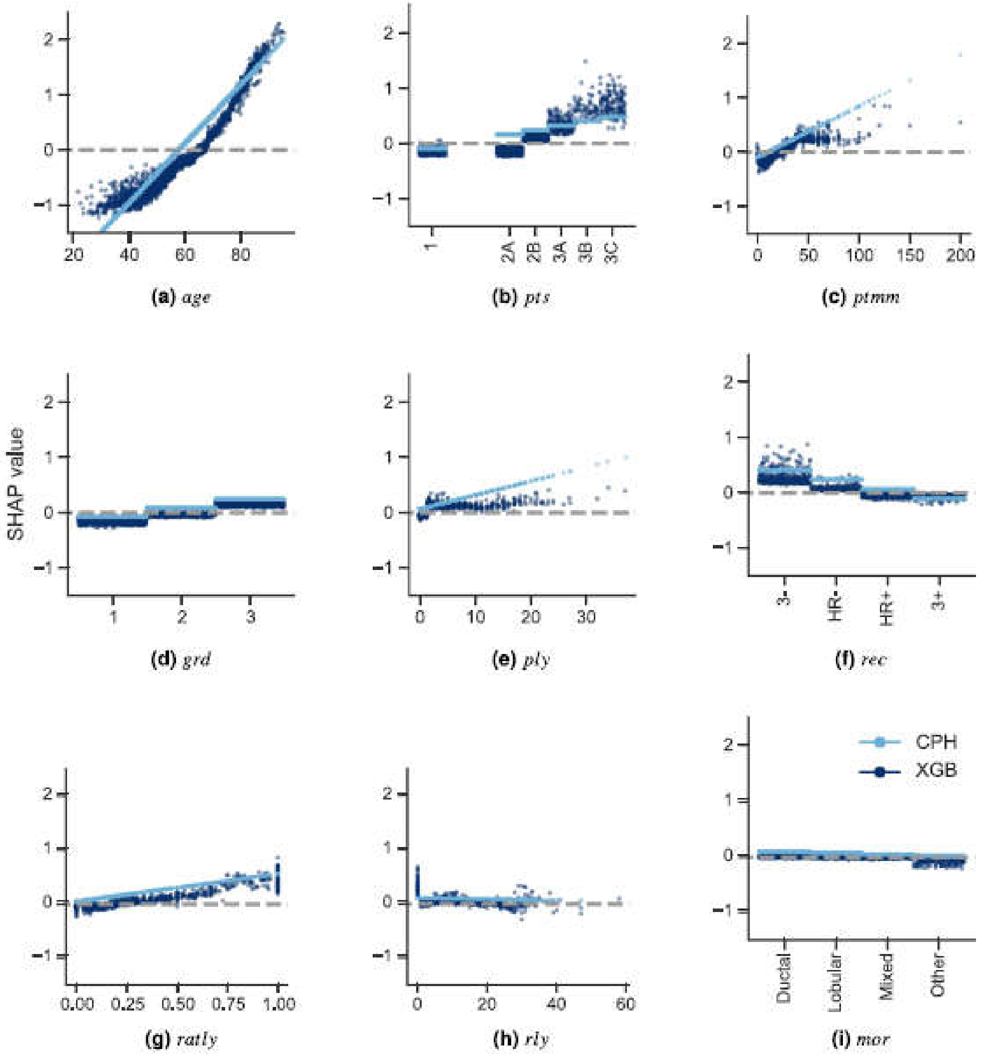
(b) XGB

Figure 2. Summary plots for SHAP values. For each feature, one point corresponds to a single patient. A point's position along the *x* axis (i.e., the actual SHAP value) represents the impact that feature had on the model's output for that specific patient. Mathematically, this corresponds to the (logarithm of the) mortality risk relative across patients (i.e., a patient with a higher SHAP value has a higher mortality risk relative to a patient with a lower SHAP value). Features are arranged along the *y* axis based on their importance, which is given by the mean of their absolute Shapley values. The higher the feature is positioned in the plot, the more important it is for the model.



(a) *age*

(b) *pts*



(a) *age* **(b)** *pts* **(c)** *ptmm*

(d) *grd* **(e)** *ply* **(f)** *rec*

(g) *ratly* **(h)** *rly* **(i)** *mor*

Figure 3. SHAP feature dependence plots. In the case of categorical variables, artificial jitter was added along the *x* axis to better show the density of the points. The scale of the *y* axis is the same for all plots in order to give a proper feeling of the magnitudes of the SHAP values for each feature (and therefore of their impact on the models' output). In the case of the XGB model, the dispersion for each possible feature value along the *y* axis is due to interaction effects (which the CPH model is unable to capture).

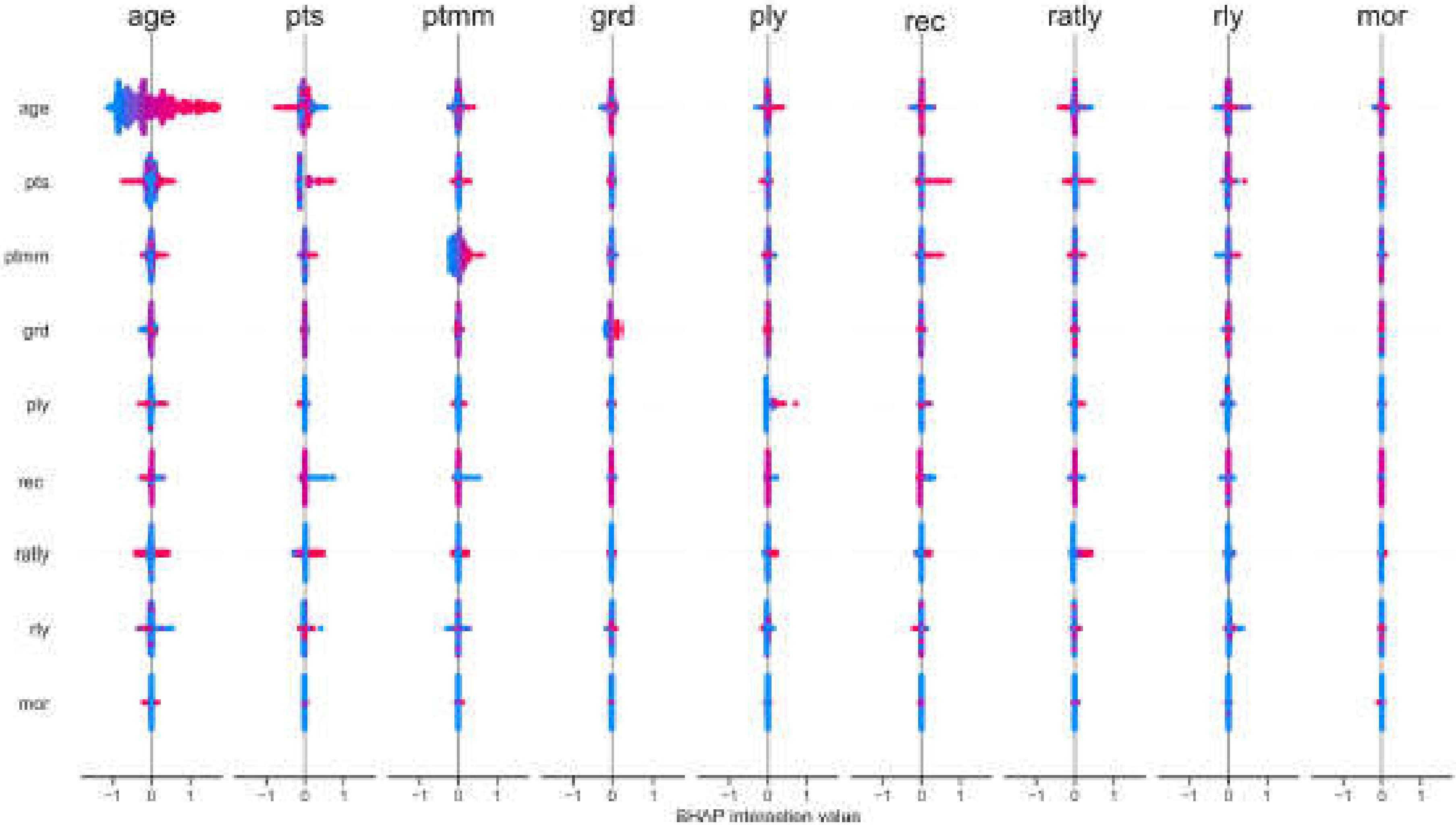
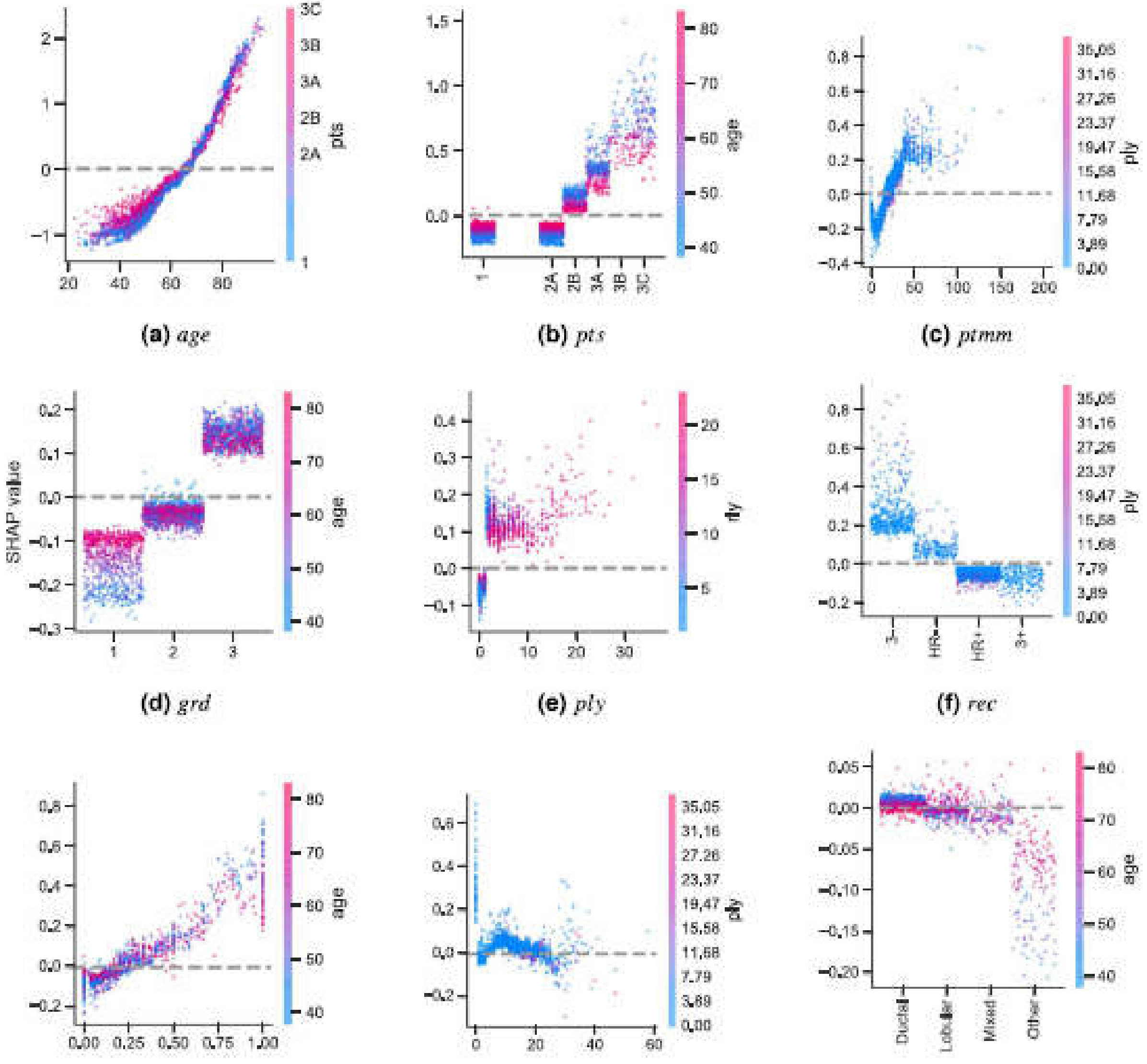


Figure 4. SHAP interaction values. The main effect of each feature is shown in the diagonal, while interaction effects are shown off-diagonal.

Figure 5. SHAP feature dependence plots of the XGB model showing the largest interaction effect for each feature. In the case of categorical variables, artificial jitter was added along the x axis to better show the density of the points. In this case, the scale of the y axis is not the same for all plots in order to better appreciate the interaction effects.



(a) *age*

(b) *pts*

(c) *ptmm*

(d) *grd*

(e) *ply*

(f) *rec*

(g) *ratly*

(h) *rly*

(i) *mor*

Conclusion

- They successfully provide the SHAP framework along with proofs and experiments showing that these values are desirable.
- The SHAP framework identifies the class of additive feature importance methods (which includes six previous methods) and allows more model interpretable and also interaction effects.

Thank you