EUCA: the End-User-Centered Explainable AI Prototyping Framework
Supplementary Material S1

1 METHOD OF DEVELOPING END-USER-FRIENDLY EXPLANATORY FORMS

We searched for XAI technique papers using "explainable/interpretable/transparency/black box" + "AI/machine learning/deep learning" in Google Scholar, IEEE Xplore Digital library, ACM Digital library, arXiv.org, and excluded works that did not evaluate the proposed method.

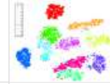
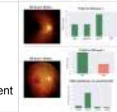
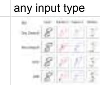
For the papers included, we performed open-coding on the type of their output explanatory information, and judged whether such information requires technical knowledge to understand. We repeated the process until information "saturated", i.e.: no new explanatory forms were identified.

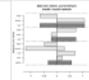
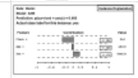
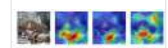
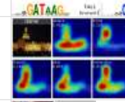
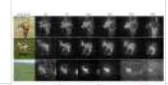
The codes revealed 12 primary types of explanatory information: feature attribution, feature shape, feature interaction, concept; decision tree, rule, counterfactual rule; instance, counterfactual instance, prototype, similar example, and clustering. Using the affinity diagram process, we grouped them into three major categories: explaining based on features, examples, and rules. They serendipitously correspond to the learned representation of a machine learning model at the feature level, instance level, and decision boundary level. We also add input, output, performance, and dataset to the explanatory forms as necessary supplementary information to make the explanation more complete.

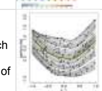
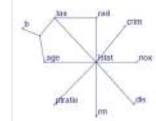
2 LIST OF REVIEWED LITERATURE

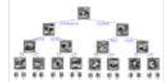
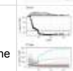
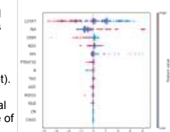
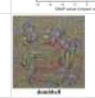
The next few pages list the reviewed XAI technical literature.

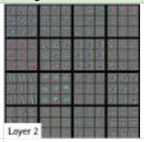
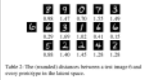
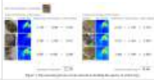
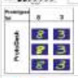
Author's address:

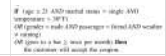
			XAI algorithm					Visual Vocabularies (Explanatory representation format class)				Local vs. global		Who	
	Algorithm Name	Paper bibiliography	Things needed to get the explanatory model (eg: model parameters, training data)	Original model (model-agnostic vs. -specific; post-hoc vs. intrinsic)	Method	XAI model output	Explanatory Information Classification	Data type	Encoding method	Vis figures	Evaluation of XAI method	Local	Global	Developers	End-users
1	t-SNE	Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579–2605. Retrieved from http://www.jmlr.org/papers/v9/vandermaaten08a.html	input data, or high-dimensional feature space	model-agnostic	non-linear transformation of high-dimensional space to 2D visualization	2D visualization	clustering	data point as clusters	dimensional reduction		visual inspection; multiple dataset comparison with other methods	☐	☑	☑	☑
2	UMAP	McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Retrieved from http://arxiv.org/abs/1802.03426	input data, or high-dimensional feature space	model-agnostic	non-linear transformation of high-dimensional space to 2D visualization	2D visualization	clustering	data point as clusters	dimensional reduction		visual inspection; computation comparison with other methods (runtime, scaliblity with embedding space, sample points)	☐	☑	☑	☑
3	iBCM	Kim, B., Glassman, E., Johnson, B., & Shah, J. (2015). iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction. Retrieved from www.csail.mit.edu	cluster label, likelihood of prototypes and subspaces	clustering method	interactive bayesian case model, user-defined clustering	user-defined clustering	clustering; prototype	prototype	show prototype and its features highlighted		user study, real-world implementation	☑	☑	☑	☑
4	TCAV	*Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (n.d.). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Retrieved from https://arxiv.org/pdf/1711.11279.pdf*	user defined examples containing pos/neg concepts; query images	CNN; classification	get the decision boundaries and its perpendicular vector as the CAV; the directional derivitative of a class training image is the TCAV	concept activation vector (showing the global class concept); measured as TCAV score (0-1)	concept	catagorical concepts, each quantified [0,1]	bar chart comparing different concepts;		simulation experiment; user test w/ lay person and doctors	☑	☑	☑	☑
5	TCAV	Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., … Cor-Rado, G. S. (n.d.). Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making, 14. https://doi.org/10.1145/3290605.3300234		CNN; image retrieval	A application using TCAV and CBIR for medical decision support		concept	catagorical concepts, each quantified [0,1]	a slider bar to control the degree of concept		mixed method user study w/ pathologist	☑	☐	☐	☑
6	network dissection	Bau, D., Zhou, B., Khosla, A., Oliva, A., & Csail, A. T. (n.d.). Network Dissection: Quantifying Interpretability of Deep Visual Representations. Retrieved from http://netdissect.csail.mit.edu	dataset with segmentation map; model with parameters	CNN; post-hoc	quantify the interpretability by aligning units in CNN with semantic concepts (segmentation)	score the semantics (ofobjects, parts, scenes, textures, materials, and colors) of hidden units at each intermediate convolutional layer. more for network analysis	concept	concept quantification	showing semantic concepts for individual units, and the layers in total.		quantify the interpretability among layers and networks	☐	☑	☑	☐
7	net2vec	Babiker, H. K. B., & Goebel, R. (2017). An Introduction to Deep Visual Explanation. Retrieved from http://arxiv.org/abs/1711.09482	training images; model parameters	post-hoc; CNN	study what information is captured by combinations (rather than individual) of neural network filters; formulate concept vectors as embeddings. theoretical analysis work, not explicitly for explanation	best filter for concept; and their learned weights (as concept embeddings)	concept	filters in CNN, and their weights	visualize the fillters of concepts, and their combined weights		quantify the filter-concept overlap w/ gt segmentation IoU	☐	☑	☑	☐
8	obj detector emerge	Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object Detectors Emerge in Deep Scene CNNs. Retrieved from http://arxiv.org/abs/1412.6856	CNN parameters; dataset w/ segmentation map to show accuracy	post-hoc; CNN; classification	visualize the unit in NN by projecting the receptive field, minimal image representations.	mask overlay on multiple input image showing the area the unit detects	concept; feature attribute		showing example images w/ masks receptive field of detect area		compare receptive field object detection w/ gt segmentation	☐	☑	☑	☑
9	Comparison-Based Inverse Classification	Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2018). Comparison-Based Inverse Classification for Interpretability in Machine Learning. In J. Medina, M. Ojeda-Aciego, J. L. Verdegay, D. A. Pelta, I. P. Cabrera, B. Bouchon-Meunier, & R. R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations (pp. 100–111). Cham: Springer International Publishing.ational Publishing. https://doi.org/10.1007/978-3-319-91479-4	input-output pairs	agnostic; classification	growing sphere: The method first draws a sphere around the point of interest, samples points within that sphere, checks whether one of the sampled points yields the desired prediction, contracts or expands the sphere accordingly until a (sparse) counterfactual is found and finally returned. They also define a loss function that favors counterfactuals with as few changes in the feature values as possible.	changed feature and its value w.r.t to the query instance	counterfactual instance; counterfactual	featurs and its new changed values, counterfactual prediction, query instance	show the instance if it's interpretable (image, text, tabular not too large) and the what-if changes in the featurs, and the counterfactual prediction		functional eval; case study	☑	☐	☑	☑
10	CNN to DT	Zhang, Q., Yang, Y., Ma, H., & Wu, Y. N. (2018). Interpreting CNNs via Decision Trees. Retrieved from http://arxiv.org/abs/1802.00121	intrinsic explanable model	intrinsic	semantic and quantatitive explanation. decomposes feature representations in high conv-layers of the CNN into elementary concepts of object parts in the decision tree. The decision tree tells people which object parts activate which filters for the prediction and how much they contribute to the prediction score.	decision tree	decision tree	semantic part outlined in the input image; the decision tree	node-link tree, show examples for the leaf		metrics (errors of object-part contributions, fitness of contribution distributions). accuracy of decision tree	☐	☑	☑	☑
11	kNN	k nearest neighbors, non-parametric, generative, supervised classification algorithm	training data	intrinsic	find the k nearest neighbors for the query instance	class label and its nearest neighbors	example	raw input	show raw input and its neighbors	any input type		☑	☐	☑	☑
12	SHAP	*Lundberg, S. M., Allen, P. G., & Lee, S.-I. (n.d.). A Unified Approach to Interpreting Model Predictions. Retrieved from https://github.com/slundberg/shap*	input features (super pixel; bag of words)	agnostic or specific	additive feature importance measure unifying (LIME, DeepLIFT, Layer-wise relevance propagation; shapley value estimation); assign each feature an important value for a prediction	unclear...	feature attribute	input feature level importance score	color code the attribute, show contrast features (remove feature to change classes)		function and human test	☑	☐	☑	☑

			XAI algorithm					Visual Vocabularies (Explanatory representation format class)				Local vs. global		Who	
	Algorithm Name	Paper bibilography	Things needed to get the explainatory model (eg: model parameters, training data)	Original model (model-agnostic vs. -specific; post-hoc vs. intrinsic)	Method	XAI model output	Explanatory Information Classification	Data type	Encoding method	Vis figures	Evaluation of XAI method	Local	Global	Develo pers	End-users
13	Interpretable Classifier for Diabetic Retinopathy Disease Grading	de la Torre, J., Valls, A., & Puig, D. (2017). A Deep Learning Interpretable Classifier for Diabetic Retinopathy Disease Grading. Retrieved from http://arxiv.org/abs/1712.08107	query image	CNN; classification	decompose the score from one layer as from input and the layer constant, using deconv	a scoring system	feature attribute	input feature level importance score	feature score at each layer for each class; and pixel-wise score		function eval and visual inspection, not thoroughly.	☑	☐	☑	☑
14	LIME	Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. Retrieved from http://arxiv.org/abs/1602.04938	sampling local instances; super pixel as iamge features, and bag of words as text features	agnostic	perturbation-based, weighted sampling around the local query instance, and fit a linear model at local	pertubation-based, support what-if by modifying feature values; depending on the explain function (linear, decision-tree, rule). In the paper they use sparse linear model.	feature attribute	input feature level importance score	image mask showing important superpixel; bar chart showing important text features		simulate gt features to test fieldity; test user for trustworthiness	☑	☑	☑	☑
15	EXPLAIN	*Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining Classifications For Individual Instances. IEEE Transactions on Knowledge and Data Engineering, 20(5), 589–600. https://doi.org/10.1109/TKDE.2007.190734*	any data type	agnostic	pertubation-based, computes the influence of a feature value by observing its impact on the model's output.	information difference measure for each features	feature attribute	neg/postive important score at input feature level [-1, 1]	bar chart (-1,1) for each features		simulation experiment for fieldity	☑	☑	☑	☑
16	IME/SHAP (Shapley Additive Explanations)	*Erikštrumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. Jmlr '10, 11, 20. Retrieved from http://www.ailab.si/orange/datasets.psp.*	input features	agnostic	pertubation-based, capture interactions between features. to reduce the computation, use game theory to approximate. generate global feature importance via game theory	feature attribute	feature attribute	neg/postive important score at input feature level [-1, 1]	bar chart pos/neg (-1,1) for each features		functional eval (fieldity, run time); qual (show explain expamples)	☑	☐	☑	☑
17	RISE	Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. Retrieved from http://arxiv.org/abs/1806.07421	input-ouput pairs; input is sampled using random masks	agnostic	pertubation-based; probe the black-box model by sub-sample the input by using random masks, and use the output as weights for the masked input	important map	feature attribute	input feature level importance score	saliency map		functional eval (insertions, deletion, pointing game accuracy)	☑	☐	☑	☑
18	Learning Global Additive Explanations	Tan, S., Caruana, R., Hooker, G., Koch, P., & Gordo, A. (2019). Learning Global Additive Explanations of Black-Box Models. https://doi.org/10.1145/nnnnnnn.nnnnnnn	input-output pairs; input features, need to be semantic meaningful so that users can interprete	agnostic	distill a student global addictive model from original teacher model. create explanation by examining the individual featuer shape w.r.t output plot.	feature shapes of a base func describes the relationship between featreus and predictions.	feature attribute	feature shape (from a base func) ploting the relationship between a feature and the output (may be non-linear)	visualize the feature shape wrt prediction (since each feature is addictive relationship with prediction); vis is suitable for ML experts, not very interpretable for end users. Need to adopt to simpler visualization.		functional eval (designing ground-truth explanations); user study with ML experts (time, capture gt features, demand, catch data error)	☐	☑	☑	☑
19	GA2M (Generalized Addictive Models plus Interactions)	Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (n.d.). Accurate Intelligible Models with Pairwise Interactions. Retrieved from http://www.cs.cornell.edu/~yinlou/papers/lou-kdd13.pdf	input-output pairs	agnostic	based on GAM (generalized addictive model) with added interaction terms of two features	GAM and important paired feature interactions	feature attribute	feature shape, paired feature interaction	line plot for feature shape, 2D heatmap for feature interaction		fidelity, case study showing the visualization	☐	☑	☑	☑
20	LRP (layer-wise relevance propagation)	Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10, e0130140 (2015).	model, weights, activation	neural network, post-hoc	it identifies important pixels by running a backward pass. The backward pass is a conservative relevance redistribution procedure, where neurons that contribute the most to the higher-layer receive most relevance from it.	pixel-level feature importance score	feature attribute	feature importance score	color code on top of the input image		visual inspection; flipping experiment	☑	☐	☑	☑
21	DeepLIFT	Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. Retrieved from http://arxiv.org/abs/1704.02685	model, activation, weights	neural network, post-hoc	compares the activation of each neuron to its 'reference activation' and assigns contribution scores according to the difference	pixel-level feature importance score	feature attribute	feature importance score	color code on top of the input image; code the importance using size on DNA data		ablation test on pixel for importance score; visual inspection	☑	☐	☑	☑
22	CAM	Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (n.d.). Learning Deep Features for Discriminative Localization. Retrieved from http://cnnlocalization.csail.mit.edu	model parameters; query image	CNN with GAP layer	weighted sum of activation maps; the weights are from GAP(global average pooling) layer	pixel-level importantce score	feature attribute	pixel-level importantce score	color coded the importance score on spatial input data		accu, localization ability, visually show the results	☑	☐	☑	☑
23	Grad-CAM & Guided Grad-CAM	Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Retrieved from http://arxiv.org/abs/1610.02391	model parameters; query image	post-hoc; CNN model family	weighted sum of activation maps, the weights are from the graidents of output w.r.t the actv maps	pixel-level importantce score	feature attribute	pixel-level importantce score; also support counterfactual explanations, by negating the gradient of target class	color coded the importance score on spatial input data (not limited to images)		user study for class discrimination, trust. analyze failure modes adversarial noise, bias.	☑	☐	☑	☑
24	SmoothGrad	Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: removing noise by adding noise. Retrieved from http://arxiv.org/abs/1706.03825	sample on the query image by adding noise; trained model	CNN; post-hoc	sample similar images by adding noise to the image, then take the average of the resulting sensitivity maps	saliency map	feature attribute	pixel-level importantce score	visualize saliency map; also visualize the difference of saliency map for top two class predictions, as a contrast explanation (or any sensitive analysis/feature attribute based method can do so), but not very intuitive (may need vis design)		visual inspection, compare w/ other grad based methods	☑	☐	☑	☑
25	PattenNet and PatternAttribute	Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., & Dähne, S. (2017). Learning how to explain neural networks: PatternNet and PatternAttribution. Retrieved from http://arxiv.org/abs/1705.05598	model parameters; input and its target output	post-hoc	disentangle the signal and weights that forms the predictions	feature attribute	feature attribute	feature-level importance score	color coded the importance score on spatial input data (not limited to images)		signal estimator quality measure; image degradation experiment; visual inspect with other methods	☑	☐	☑	☑

			XAI algorithm					Visual Vocabularies (Explanatory representation format class)				Local vs. global		Who	
	Algorithm Name	Paper bibilography	Things needed to get the explanatory model (eg: model parameters, training data)	Original model (model-agnostic vs. -specific; post-hoc vs. intrinsic)	Method	XAI model output	Explanatory Information Classification	Data type	Encoding method	Vis figures	Evaluation of XAI method	Local	Global	Develo pers	End-users
26	right for the right reasons	Ross, A., Hughes, M. C., & Doshi-Velez, F. (n.d.). Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. Retrieved from https://github.com/dtak/rrr.	input	post-hoc	align gradient-based method with pertubation-based method, since pertubation methods are computational expensive; input gradient explanations match state of the art sample-based explanations; optimize the classifier to learn alternative explanations.	feature importance	feature attribute		feature positive/negative attribute		visual comparion w/ LIME baseline	☑	☐	☑	☑
27	Distill-and-Compare	Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. https://doi.org/10.1145/3278721.3278725	audit data (not necessaryly training data); gt; black-box model	agnostic	compare the student model trained with distillation to a second un-distilled transparent model trained on ground-truth outcomes, and use differences between the two models to gain insight into the black-box model	use iGAM as transparent model in the paper; feature contributions	feature attribute		in the form of GAM or tree (depending on the explanatory model used)		fidelity of the mimic model	☐	☑	☑	☑
28	deep visual explanation	Babiker, H. K. B., & Goebel, R. (2017). An Introduction to Deep Visual Explanation. Retrieved from http://arxiv.org/abs/1711.09482	model; query image	CNN	transform the activation map in Fourier domain, and convert back to get the saliency map	saliency map	feature attribute		saliency map		visual inspect w/ other saliency map method	☑	☐	☑	☑
29	Prospector	Krause, J., Perer, A., & Ng, K. (n.d.). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. https://doi.org/10.1145/2858036.2858529	input-output pairs	agnostic	an interactive visual analytic system based on partial dependence plot	partical dependence of features for global and individual explanation	feature attribute	feature shape	color bar; line chart		case study on predicting diabetes on EHR w/ data scientists	☑	☑	☑	☑
30	Individual conditional expectation plot (ICE)	Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2013). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. Retrieved from http://arxiv.org/abs/1309.6392	input-output pairs	agnostic	based on the partial dependence plot, and graph the functional relationship between the predicted response and the feature for individual observations. It suggests where and to what extent heterogeneities might exist.	feature shape for individual data point	feature attribute	feature shape for individual data point	line and scatter plot for each individual data point, showing the heterogeneity of the effects		visual test for addictivity; simulated and real data inspection	☑	☑	☑	☑
31	VIN (Variable interaction network)	G. Hooker. Discovering additive structure in black box functions. In Pro- ceedings ofthe tenth ACM SIGKDD international conference on Knowl- edge discovery and data mining, pp. 575–580. ACM, 2004	input-output pairs	agnostic	features are displayed in a stylized network graph in which connections indicate the presence of an interaction. This method is notable for its ability to efficiently identify interactions including 3 or more terms. The interactions are identified by an algorithm that uses a permutation method similar to feature importance scores [6] to identify features whose effect changes in the presence or absence of a potential interactor feature. The algorithm then cleverly prunes the search space by using the property that an interaction effect can only exist if all the lower-order effects that involve its feature also exist	interaction strength	feature attribute	variable interaction network as a graph; this work can extend the vis in feature attribute by visualizing the interactions of features as graph	node-link undirected graph		show case study	☐	☑	☑	☑
32	Mind the Gap	Kim, B., Shah, J. A., & Doshi-Velez, F. (2015). Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. Retrieved from https://papers.nips.cc/paper/5957-mind-the-gap-a-generative-approach-to-interpretable-feature-selection-and-extraction	intrinsic	intrinsic generative model	graphical model for feature selection	distinguishable feature dimensions, and their clusters	feature attribute	feature value	visually show the distinguishable features		baseline eval; user study	☐	☑	☑	☑
33	RETAIN	Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016). RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. Retrieved from http://arxiv.org/abs/1608.05745	model, trainining data	intrinsic interpretable RNN model	use attention model to detect influential past visits and significant clinical variables within those visits	feature contribution in EHR	feature attribute	feature contribute	visualize the feature contribution on a time scale		model performance; visual inspection	☑	☐	☑	☑
34	Integrated Gradient	Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. Retrieved from http://arxiv.org/abs/1703.01365	model, gradient	CNN; post-hoc	combines the Implementation Invariance of Gradients along with the Sensitivity of techniques like LRP or DeepLift	pixel-level feature importance score	feature attribute	feature importance score	color coded the importance score on spatial input data		visual inspection; heatmap showing the feature correlation between the language translation model	☑	☐	☑	☑
35	PDP	Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232. Retrieved from http://www.jstor.org.proxy.lib.sfu.ca/stable/2699986	input-output pairs	agnostic	get the marginal effect of features (1 or 2) on the prediction	feature value w.r.t prediction, feature shape	feature attribute	feature shape	line or surface plot		multiple dataset visual inspection	☐	☑	☑	☑
36	Interpretable CNN	Zhang, Q., Wu, Y. N., & Zhu, S.-C. (2017). Interpretable Convolutional Neural Networks. Retrieved from http://arxiv.org/abs/1710.00935	intrinsic explanable model	intrinsic	the loss function make the filers in the deep layer CNN represent the specific object part	visualize the filter as object detector	feature attribute	input image with mask showing the receptie field of the filters	image mask		classification accuracy; location stability; visual inspection	☑	☑	☑	☑

			XAI algorithm					Visual Vocabularies (Explanatory representation format class)				Local vs. global		Who	
	Algorithm Name	Paper bibilography	Things needed to get the explainatory model (eg: model parameters, training data)	Original model (model-agnostic vs. -specific; post-hoc vs. intrinsic)	Method	XAI model output	Explanatory Information Classification	Data type	Encoding method	Vis figures	Evaluation of XAI method	Local	Global	Develo pers	End-users
37	distillation	Watanabe, C., Hiramatsu, K., & Kashino, K. (2018). Knowledge Discovery from Layered Neural Networks based on Non-negative Task Decomposition. Retrieved from https://arxiv.org/pdf/1805.07137.pdf Barrett, D. G. T., Morcos, A. S., & Macke, J. H. (2018). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? Retrieved from https://arxiv.org/pdf/1810.13373.pdf Che, Z., Purushotham, S., Khemani, R., & Liu, Y. (n.d.). Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. Retrieved from https://arxiv.org/pdf/1512.03542.pdf Distilling a Neural Network Into a Soft Decision Tree.	trained model, training data	post-hoc	teach an interpretable model by learning from black-box model, using its output as soft labels	as the format of interpretable model: linear, decision tree/rule	feature attribute; decision		depends on the form of interpretable model		compare the student model performance with teacher model	☐	☑	☑	☑
38	Gamut	Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19 (pp. 1–13). New York, New York, USA: ACM Press. https://doi.org/10.1145/3290605.3300809	input-output pairs	agnostic	visual analytic system based on GAM curves	partical dependence of features for global and individual explanation	feature attribute; linear	feature importance score	mainly line plot for features, also support instance explanation and user defined grouping		participatory design; thorough user study	☑	☑	☐	☑
39	**VINE**	Britton, M. (2019). VINE: Visualizing Statistical Interactions in Black Box Models. Retrieved from http://arxiv.org/abs/1904.00561	input-output pairs	agnostic	reginal explanations, i.e. algorithm capture a subset of data that share a common behavior (like unsupervised clustering), and describe the common behaviro. capture the feature interaction which is a weakness in partial dependence plot	VINE curve, showing the PDP/IDE plot, and the decompositions from regional explanations	feature attribute; linear	feature values, and interaction strength (another dimension to be added to the feature attribute class)	encode the PDP as line chart; encode the individual line chart on a 2D plot; also plot the PDP as 2-D feature heatmap and contour plots. (Note that PDPs (and other plots in this family) can be presented with the standard scale (in which the Y-axis is read as the predicted value) or as a centered PDPs (and other plots in this family) can be presented with the standard scale (in which the Y-axis is read as the predicted value) or as a centered PDP (in which case the Y-axis is read as the change from the average prediction)		compare to random clustering baseline and statsitical methods	☐	☑	☑	☑
40	Visualizing the Feature Importance	Casalicchio, G., Molnar, C., & Bischl, B. (2018). Visualizing the Feature Importance for Black Box Models. Retrieved from http://arxiv.org/abs/1804.06620	input-output pairs (black-box)	agnostic	pertubation/sampling-based using Monte-Carlo to measure feature importance on individual data	local feature importance score	feature attribute; linear	local and global importance score	partial importance (PI); individual conditional importance (ICI) plots as line plot		simulation experiment; real data	☑	☑	☑	☑
41	Tree SHAP	Lundberg, S. M., Erion, G. G., & Lee, S.-I. (n.d.). Consistent Individualized Feature Attribution for Tree Ensembles. Retrieved from http://github.com/slundberg/shap	input-output pairs; trees	tree ensembles; specific	estimate SHAP values and interaction for tree ensembles	SHAP values (individualized feature attribute); cluster samples by explnation similarity (of different feature combinations/interactions)	feature attribute; linear	data subset clustering; global feature importance	data subset clustering; partial dependence plot (bar chart representing global feature importance); SHAP summary plots (plot each individual dot on the global feature attribute plot, dot is color coded by the feature value); SHAP dependence plot (plot invidicual data in the partial dependence plot). An aggregation of local explanation to form a global explanation is also the role of visual analytics.		AUC; user study agreement w/ human	☑	☑	☑	☑
42	sensitivity analysis & class prototype	Simonyan, K., Vedaldi, A., & Zisserman, A. (n.d.). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Retrieved from http://code.google.com/p/cuda-convnet/	model, weight, gradient	CNN; post-hoc	gradient-based saliency map; optimization to find the class prototype	saliency map; class prototypical image	feature attribute; prototype	feature importance; prototype image	color coded the importance score on spatial input data		visual inspection	☑	☑	☑	☑
43	GuidedBackProp	Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for Simplicity: The All Convolutional Net. In ICLR workshop. Retrieved from http://arxiv.org/abs/1412.6806	model; gradient	post-hoc; CNN model family	combine deconvolution and gradient back prop to get sparse feature attribute	pixel-level importantce score; filter visualization	feature attribute; protytpe	pixel-level importantce score; filter visualized as object detector	color coded the importance score on spatial input data; filter visualization		visual inspection	☑	☐	☑	☑

			XAI algorithm					Visual Vocabularies (Explanatory representation format class)				Local vs. global		Who	
	Algorithm Name	Paper bibilography	Things needed to get the explainatory model (eg: model parameters, training data)	Original model (model-agnostic vs. -specific; post-hoc vs. intrinsic)	Method	XAI model output	Explanatory Information Classification	Data type	Encoding method	Vis figures	Evaluation of XAI method	Local	Global	Developers	End-users
44	Deconv	Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8689 LNCS, pp. 818–833). Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53	model; gradient	post-hoc; CNN model family	use deconvolution operation to backprop the decision to input space	pixel-level importantce score; filter visualization	feature attribute; protytpe	pixel-level importantce score; filter visualized as object detector	color coded the importance score on spatial input data; filter visualization		occulusion test, visual inspection	☑	☑	☑	☑
45	Wachter's counterfactual explanation	Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Retrieved from http://arxiv.org/abs/1711.00399	input-output pairs	agnostic	minimize a counterfactual instance as close as the query instance such that its prediction is the counterfactual prediction	unconditional counterfactual explanations	instance; counterfactual	counterfactual instance (with the most changed features), and counterfactual prediction	text to show the tabular instance feature and its prediction		unclear...	☑	☐	☐	☑
46	Prototype case-based reasoning	*Li, O., Liu, H., Chen, C., & Rudin, C. (n.d.). Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. Retrieved from https://arxiv.org/pdf/1710.04806.pdf*	training dataset to train the XAI model; query image for similarity measure	intrinsic; VAE; classification	a prototype layer; cost func minimize the prototype vector to be close to the training set; visualize the prototype vector using decoder	learned class prototypes	prototype		showing prototypical examples as what the NN learned; similarity distance between query and prototyeps		visual inspect the prototypes, similarity distance of query images to prototypes	☑	☑	☑	☑
47	This looks like that	*Chen, C., Li, O., Tao, C., Barnett, A., & Rudin, C. (n.d.). This Looks Like That: Deep Learning for Interpretable Image Recognition. Retrieved from https://arxiv.org/pdf/1806.10574.pdf*	training dataset to train the XAI model	intrinsic; CNN; classification	a protoytpe layer in CNN replace conv opertaion with squared L2 distance computation to training patches (as prototype filter); final prediction is the linear combination of prototype layer; add seperation and cluster cost.	prototypes are prototypical parts of images	prototype		activation map of prototype + similarity score + total points for class; complex reasoning process		visual inspection of explanatory, and tSNE for visuallizing latent space learned by the model; accu	☑	☐	☐	☑
48	Bayesian case model	Kim, B., Rudin, C., & Shah, J. (2014). The Bayesian case model: a generative approach for case-based reasoning and prototype classification. Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. MIT Press. Retrieved from https://dl-acm-org.proxy.lib.sfu.ca/citation.cfm?id=2969045	intrinsic	intrinsic model	perform joint inference on cluster labels, prototypes and important features to learn prototype	prototype and subspace	prototype	prototype and subspace	show prototype and subspace		user study; visual inspection	☐	☑	☑	☑
49	ProtoDash	Gurumoorthy, K. S., Dhurandhar, A., & Cecchi, G. (2017). ProtoDash: Fast Interpretable Prototype Selection. Retrieved from http://arxiv.org/abs/1707.01212	input dataset	clustering method	prototype identification with weights, based on learn to criticise	weighted prototypes	prototype	prototype	show prototype		visual inspection; user study	☐	☑	☑	☑
50	attention-based prototypical learning	Arik, S. O., & Pfister, T. (2019). Attention-Based Prototypical Learning. Retrieved from http://arxiv.org/abs/1902.06292	neural network with attention module	neural network; post-hoc	utilizes an attention mechanism that relates the encoded representations to determine the prototype	class prototype and its weights	prototype	prototype	prototype		visual inspection of image and text prototypes; robust to label noise, sparse explanation	☐	☑	☑	☑
51	k-Medoids	KAUFMANN, & L. (1987). Clustering by Means of medoids. Proc. Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, 1987, 405–416. Retrieved from https://ci.nii.ac.jp/naid/10027761751/	training data	intrinsic, finding prototypes		k-medoids	prototype; clustering	raw input, medoids	show input data and prototypes	any input type		☑	☑	☑	☑
52	MMD-critic	Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. Retrieved from https://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability	training data (to find the prototype and critism)	debug for the model, input data distribution	nearest prototype model': get representative instances (prototypes and critism) to debug the model. using greedy search to find prototypes which represents the dataset, and crititism (outliers) which not represented by the prototype. compares the distribution (measured by witness function using RBF kernel) of the data and the distribution of the selected prototypes	get the model's predictions for prototypes and critisms, and debug based on it. understand complex data distributions	prototype; clustering	input data instance	show input data		uesr study show users have better performance using prototypes and critisms than random images	☑	☑	☑	☑
53	RuleMatrix	Ming, Y., Qu, H., & Bertini, E. (n.d.). RuleMatrix: Visualizing and Understanding Classifiers with Rules. Retrieved from https://arxiv.org/pdf/1807.06228.pdf	input-output pairs	agnostic	pedagogical learning, student rule use the labels from the teacher model; rule learning based on Scalable Bayesian Rule Lists; rule filter to make the explanation selective	rules	rule	data flow; rules (feature, rule support and fiedlity); data distribute to indicate the rule	matrix row - rule, col - feature, grid - feature distribute. show data flow as the order of the rule; support info show the right/wrong ratio, fidelity, evidence. User can interact to filter the rules.		user case and user study, no evalaution on the rule indcution algorithm	☑	☑	☐	☑
54	Anchor	Ribeiro, M. T., Singh, S., & Guestrin, C. (n.d.). Anchors: High-Precision Model-Agnostic Explanations. Retrieved from www.aaai.org	perturbation distributions and a black box model	agnostic	rule finding algorithms not assume a dataset prior	An anchor explanation is a rule that sufficiently "anchors" the prediction locally – such that changes to the rest of the feature values of the instance do not matter.	rule	anchored feature for an query instance, and precision and coverage https://github.com/marcotcr/anchor	if then rule list		simulation experiment; user study	☑	☐	☑	☑
55	Bayesian Rule Lists	Letham, Benjamin, et al. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model." The Annals of Applied Statistics 9.3 (2015): 1350-1371.	training data to train the interpretable model	classificaition; intrinsic	produce decision lists using generative model, producing a posterior distribution over if then rules; employs a novel prior structure to encourage sparsity.	trained interpretable model of rule list, for medical scoring and grading	rule	rules and predicted class probabilities and (CI)	if else text description list		AUC of the model	☑	☑	☐	☑

			XAI algorithm					Visual Vocabularies (Explanatory representation format class)				Local vs. global		Who	
	Algorithm Name	Paper bibilography	Things needed to get the explainatory model (eg: model parameters, training data)	Original model (model-agnostic vs. -specific; post-hoc vs. intrinsic)	Method	XAI model output	Explanatory Information Classification	Data type	Encoding method	Vis figures	Evaluation of XAI method	Local	Global	Developers	End-users
56	Scalable Bayesian Rule Lists	Yang, H., Rudin, C., & Seltzer, M. (2016). Scalable Bayesian Rule Lists. Retrieved from http://arxiv.org/abs/1602.08610	training data to train the interpretable model	classificaition; intrinsic	built upon a pre-mined rules; global optimization (instead of DT of greedy optimize) defining a distribution of decision lists with prior distributions for the length of conditions (preferably shorter rules) and the number of rules (preferably a shorter list).	trained interpretable model of rule list	rule	rules	if else text description list		AUC and runtime of the model	☑	☑	☐	☑
57	Bayesian Rule Sets	Wang, T., Rudin, C., Velez-Doshi, F., Liu, Y., Klampfl, E., & MacNeille, P. (2016). Bayesian Rule Sets for Interpretable Classification. In 2016 IEEE 16th International Conference on Data Mining (ICDM) (pp. 1269–1274). IEEE. https://doi.org/10.1109/ICDM.2016.0171	intrinsic model; training data	intrinsic model	a Bayesian framework for learning rule set models, with prior parameters can be set by users to encourage the model to have a desirable size and shape	rule sets	rule	rules	if else text description list		test on 10 UCI dataset with other baseline interpretable models	☐	☑	☑	☑
58	Surrogate Decision Tree Visualization	Castro, F. Di, & Bertini, E. (2019). Surrogate Decision Tree Visualization Interpreting and Visualizing Black-Box Classification Models with Surrogate Decision Tree. Retrieved from http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-15.pdf	input-output pairs for train the decision tree, with training data and their soft labels (labelled by original model)	agnostic	use model distillation to train the decision tree on soft labels/	decision tree, and feature importance (quantified by Gini index)	rule	decision tree. user can select the tree depth by sliding the fidelity level	Tree: node-link; rule: tabluar		functional (fidelity, computational speed, tree complexity); user study w/ ML developers	☑	☑	☑	☑
59	LORE	Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. Retrieved from http://arxiv.org/abs/1805.10820	input-output pairs	agnostic	genetic algorithms for neighborhood generation	local explanations consists of 1) local rule and 2) counterfactual rule	rule; conterfactual	decision tree, rule list	tree or rule list		fidelity compare with other baseline method lime, anchor	☑	☐	☑	☑
60	ALE	Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. https://arxiv.org/abs/1612.08468	input-output pairs	agnostic	accumulated local effects (ALE) plots, estimate feature-prediction relationship that do not require this unreliable extrapolation with correlated predictors	feature value w.r.t prediction, feature shape	feature	feature shape	line or surface plot		simulation experiment	☐	☑	☑	☑
61	CBR (Case-Based Reasoning), CBIR (content based image retrieval)	An Introduction to Case-Based Reasoning. http://alumni.media.mit.edu/~jorkin/generals/papers/Kolodner_case_based_reasoning.pdf	input, training data, w/o model	depends	different CBR algorithms	similar examples	examples	similar examples	example			☑	☐	☑	☑
62	k-medoids	https://en.wikipedia.org/wiki/K-medoids	input, training data, w/o model	depends	chooses actual data points as centers (medoids or exemplars)	prototypical examples	examples	prototypical examples	example			☑	☑	☑	☑
63	Influential instance	Understanding Black-box Predictions via Influence Functions. https://arxiv.org/abs/1703.04730	input, model	agnostic; post-hoc	identify training points most responsible for a given prediction	prototypical examples	examples	prototypical examples	example		only functional validation	☑	☑	☑	☑
64	Pertinent Negative	Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. https://papers.nips.cc/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf	input, trained model, autoencoder	model-agnostic; post-hoc	optimize the selectet image perturbation given the loss function	counterfactua example and counterfactual feature highlighting	examples	counterfactual	example		human expert user study on three tasks	☑	☐	☑	☑
65	Counterfactual Visual Explanations	https://arxiv.org/abs/1904.07451	input, trained model	agnostic	identify img regions that switch the decision	counterfactua example and counterfactual feature highlighting	examples	counterfactual	example		quant: img edit experiment; visual inspection	☑	☐	☑	☑
66	Progression	Graph Geodesics to Find Progressively Similar Skin Lesion Images. www2.cs.sfu.ca/~hamarneh/ecopy/miccai_grail2017.pdf	input, trained model	agnostic; post-hoc	compute the geodesic/shortest path between nodes to determine a path of progressively visually similar skin lesions.	counterfactual, similar example	examples	counterfactual, similar example	example		proposed metrics to compare with proxy gt of classification labels. Compared with baseline model	☑	☐	☑	☑