Available online at **www.joac.info**

ISSN: 2278-1862

Journal of Applicable Chemistry

2023, 12 (5): 835-915

(International Peer Reviewed Journal)

New Chemistry News

$^{-}N=C=N^{-}$

New News of Chem (NNC) **ChemNewsNew (CNN)**

CNN-58--Fit (Figure Image TableScript...)BasesPart 6. xAI(Bfit) 2022-2023 Probes

Information Source	sciencedirect.com;	
S. Narasinga Rao M D **Associate Professor,** **Dept. of General Medicine,** **Government medical college,** **government general hospital,** **Srikakulam, AP, India**	**K. SomasekharaRao, Ph D** **Dept. of Chemistry,** **Acharya Nagarjuna Univ.,** **Dr. M.R.Appa Rao Campus,** **Nuzvid-521 201, India**	**R. Sambasiva Rao, Ph D** **Dept. of Chemistry,** **Andhra University,** **Visakhapatnam 530 003,** **India**
snrnaveen007@gmail.com (+91 9848136704)	sr_kaza1947@yahoo.com (+91 98 48 94 26 18)	rsr.chem@gmail.com (+91 99 85 86 01 82)

Conspectus: The tasks based on physics, chemistry and biology are modelled with empirical, theoretical or computational approaches. The I2O (Input-to-Output) transformation through the best of best models remained to be black-box approaches except in the case simple regression or trees. In the last decade, there was an upsurge to understand as deeply as possible the model employed, I/O transformation, parameter space, transformations, logic in arriving at intermediate information/ knowledge/hypothesis/ conclusions/ acceptance or rejection of advices. Further, it is used to point out explicit explanation not only to users but also to all stake holders. This protocol is indispensable and essential in health sector,

Défense, legal affairs, environmental policies, manufacure and so on.

xAI:In 2015, DARPA (USA) coined the term xAI(explainable Artificial Intelligence) and within a span of few years, it turned out into an indispensable trans-discipline in science/engineering/technology. Under the umbrella-xAI, noteworthy mathematical probes emerged enabling explanation frame for complex machine learning work-flows, deep neural nets, (vector/matrix) capsule nets etc. It altogether changed the mode of reporting of modelling output. The DNA approach for probes, software-tools, display methods follow.

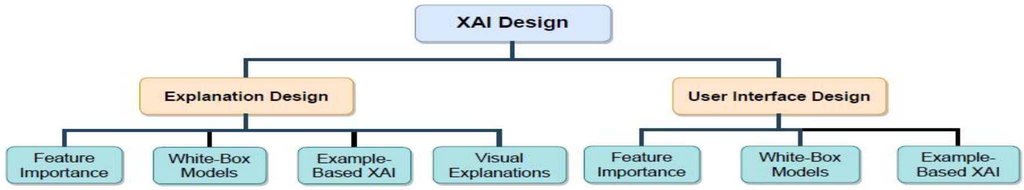
Tasks in Mathematical Language: Broad types of tasks with xAI are detection, classification, Segmentation, clustering, regression, and structure-property/ structure-function/ structure-response relationships.
Disciplines employing xAI:The applied and trans-areas employing xAI are medicine, Molecular/material science, environment, Nuclear physics, molecular genetics etc.
Data types:The different data modes used as input are tabular (numerical, categorical, binary, logical), text (words, sentences, scripts), images (2D-,3D-, RGB), point clouds, audio, graphs, videos etc.
Models: Models are broadly classified as black-box, grey-box and white-box types. They are also otherwise called as transparent and opaque.
xAI-probes: The two important categories of xAI-probes are local and global. Another division is ante-hoc and post-hoc.
Libraries: Available are AIX-360.Captum, InterpretML, Skater Ecco and XATIK
Frameworks: Are developed in Python for post-hoc XAI of DeepNNs. Typical ones are Zennit, Captum, and Nvestigate
Explanation modes:If-Then-Else, scripts, Graphical (Figures, images, videos) and multi-media (Text and graphs, video/audio)
Explanation objects: Areall layersof neurons, layers of capsules, features, processes, decisions
Explanation methods: The explanation is sometimes by simplification. But, mostly based on perturbation of input/output/parameters of model, gradient and the concept.
Feature Relevance explanation in NNs is monitored and assessed by integrated gradients, guided BP, Layerwise Relevance propagation, Graph LRP, Deep Taylor decomposition, DeepLift (Learning Important FeaTures), Concept activation, activation maximization and Prediction difference Analysis (PDA).
Some other typical xAI probes are Local interpretable model-agnostic explanations (LIME), Sub modular pick (SP)-LIME, anchor-LIME, LORE, SHAP, Shapley additive explanations, Saliency Maps, Class model visualization, LOKE, Anchors, class activation map (CAM) ,Grad-CAM, Grad-CAM++, SMOOTHGRAD, U-CAM, Eigen-CAM, DeepRed, GAM, Decision Trees, LENs and BRL.The output (Fit: Figure Image Table Script …. Bases) of typical case studies using xAI-probes during 2022Jan to 2023June are described.

Keywords:xAI, Post-hoc, ante-hoc explanations; xAI-Probes; Local interpretable model-agnostic explanations (LIME), SHAP, Layerwise Relevance propagation, Partial dependence plots,Class Activation map (CAM), Grad-CAM; Integrated gradients;Concept activation map, Heatmaps; Saliency maps;;tSNE plot; Feature Relevance explanation, Rule extraction, eXplainable/ interpretableNumerical values,Figures; Images, Tables, Scripts.

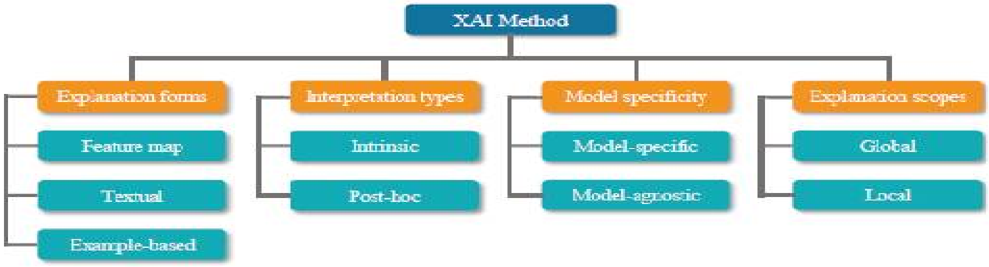
Probes of xAI methods

xAI.		2023-145

Classification of XAI design methods

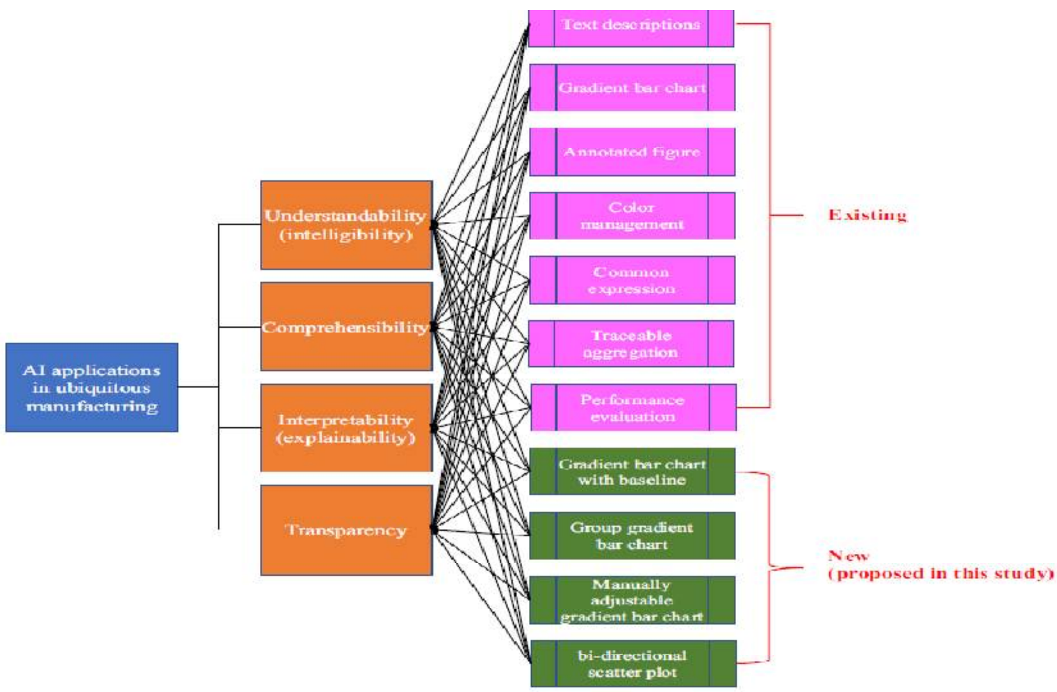


xAI.		2023-105



xAI.	3D printing	2023-148

XAI Techniques and Tools



xAI. 2023-145

XAI methods

Explanation Type	Black-Box Model	Method	Scope	Functionality
Feature Importance	Any	LIME	Local	Surrogate Model
		LORE	Local	Surrogate Model
		Anchors	Local	Surrogate Model
		Occlusion	Local	Input Perturbation
		Permutation Feature Importance	Global	Input Perturbation
		Shapley Feature Importance	Global	Game-Theory
		SHAP	Both	Game-Theory
	Neural Network	Guided Backpropagation	Local	Backpropagation
		Integrated Gradients	Local	Backpropagation
		Layerwise Relevance Propagation	Local	Backpropagation
		DeepLift	Local	Backpropagation
		Testing with Concept Activation Vectors	Global	Human Concepts
		Activation Maximization	Global	Forwardpropagation
	CNN	Deconvolution	Local	Backpropagation
		Class Activation Map	Local	Backpropagation
		Grad-CAM	Local	Backpropagation
	Transformer	Attention Flow / Attention Rollout	Local	Network Graph
		Transformer Relevance Propagation	Local	Backpropagation
White-Box Model	Any	Rule Extraction	Global	Simplification
		Tree Extraction	Global	Simplification
		Model Distillation	Global	Simplification
	CNN	Attention Network	Global	Model Adaption
	RNN	Attention Network	Global	Model Adaption
Example-Based	Any	Prototypes	Global	Example (Train Data)
		Critisisms	Global	Example (Train Data)
		Counterfactuals	Global	Fictional data point
Visual Explanations	Any	Partial Dependence Plot	Global	Marginalization
		Individual Conditional Expectation	Global	Marginalization
		Accumulated Local Effects	Global	Accumulation

xAI. 2023-150

List of XAI studies that used EBM to explain their model prediction results

Author	Objective	Subject	Application	Data type	Sig. Features	ML model	Results
Magunia et al. [63]	Identifies ICU outcome predictors in a multicenter COVID-19 cohort	1186 patients	ICU/ patient outcome	EHR/EMR	age, platelet/neutrophil ratio, D-dimer, admission by external transfer, Murray lung injury score, D-dimer level, creatinine level, SOFA score w/o GCS	EBM	Survival ACC: 64.00% AUROC: 0.810 ECMO therapy ACC: 73.00% AUROC: 0.690 Renal replacement therapy ACC: 70.00% AUROC: 0.690
Qu et al. [64]	Identify Predictors of Congenital Heart Diseases	119 CHD 239 normal	CDS	Questionnaires and clinical data	maternal coagulation function indicators, glucose levels, maternal serum UA levels	EBM	ACC: 65.00% SPE: 65.00% SEN: 74.00% AUROC: 0.760

List of XAI studies that used a rule-based system to explain their model prediction results

Author	Objective	Subject	Application	Data type	Data	ML/DL	Classifier	Technique	Results
Gidde et al. [66]	COVID-19 Detection	6 Public dataset	CDS	Image	Chest X-ray	DL	U Net, R CNN, DenseNet-201	Expert system devised by radiologists	ACC: 95.00% SPE: 97.00% SEN: 78.00% AUROC: 0.890
Mellem et al. [67]	clinical trial patient selection to retrospectively improve treatment effects in schizophrenia	95 treatment arms 102 placebo	Treatment effect study	clinical data	clinical trial data	ML	Rule-based	Bayesian Rule Lists	ACC: 71.10% AUROC: 0.740

xAI. | 2023-081

xAI Probes

	Result Type		Scope		Role	
	feature-scoring	rule-based	local	global	interpret. model	expl. method
LIME	✔		✔			✔
SHAP	✔		✔			✔
Activation Maximization	✔		✔			✔
Saliency Maps	✔		✔			✔
SP-LIME	✔			✔		✔
Class Model Visualization	✔			✔		✔
LORE		✔	✔			✔
Anchors		✔	✔			✔
DeepRED		✔	✔	✔		✔
GAM	✔			✔	✔	
Decision Trees		✔	✔	✔	✔	
BRL		✔	✔	✔	✔	
LENs		✔	✔	✔	✔	✔

xAI. | 2023-104

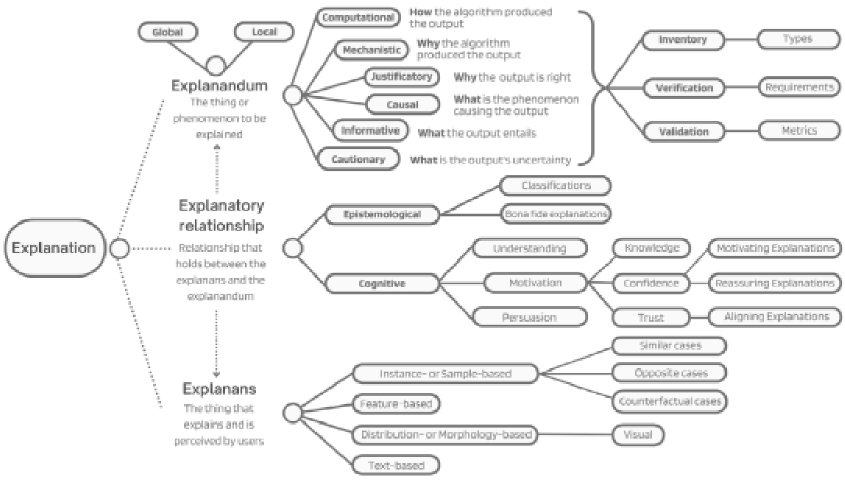
Key properties of state-of-the-art algorithms

Overview of key properties of the proposed model and other state-of-the-art algorithms. The order of appearance of the state-of-the-art algorithms in the table is the same as in the discussion in the literature review section.

property	this paper	LIME [33]	SHAP [23]	LRP [3]	NAM [1]	CHIRPS [15]	LORE [14]	MOC [7]	MAPLE [30]	DICE [26]	FACE [31]
fast	✔	✔		✔		✔	✔				✔
deterministic	✔		✔	✔							
local	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	
model-agnostic	✔	✔	✔				✔	✔		✔	✔
suitable for multi-class models	✔	✔	✔	✔	✔	✔		✔	✔	✔	✔
user-parameter-free							✔		✔		
post hoc	✔	✔	✔	✔		✔	✔	✔	✔	✔	✔
provides counterfactual explanations	✔					✔	✔	✔	✔	✔	✔
provides symbolic explanations	✔						✔				
explanations from the domain	✔			✔		✔			✔	✔	
certainty quantification	✔				✔	✔			✔		

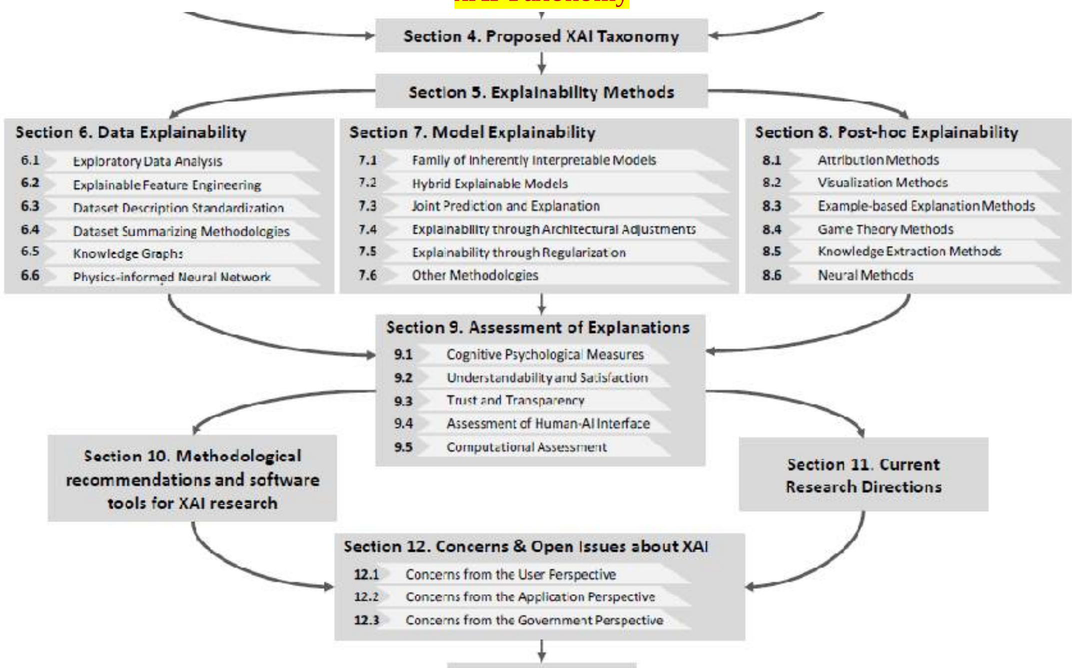
xAI. | 2023-116

Topology of explanations in xAI



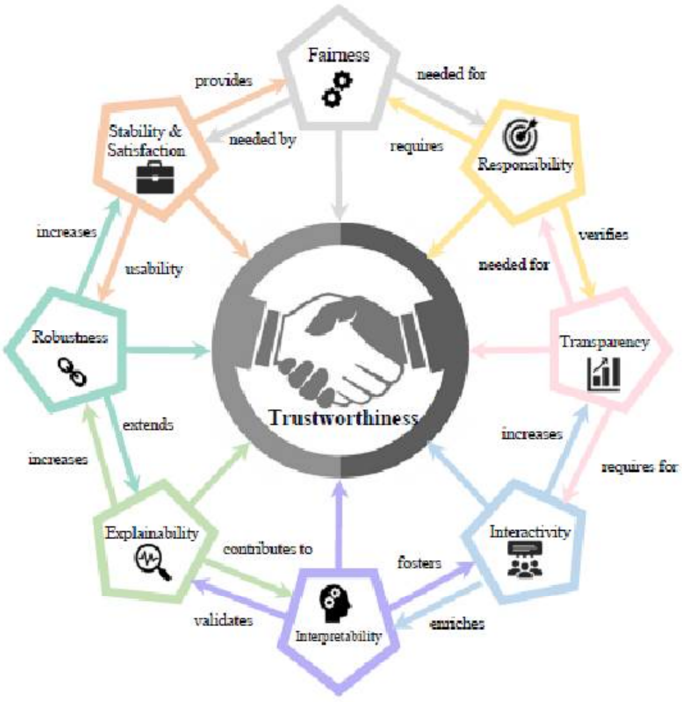
xAI. | 2023-030

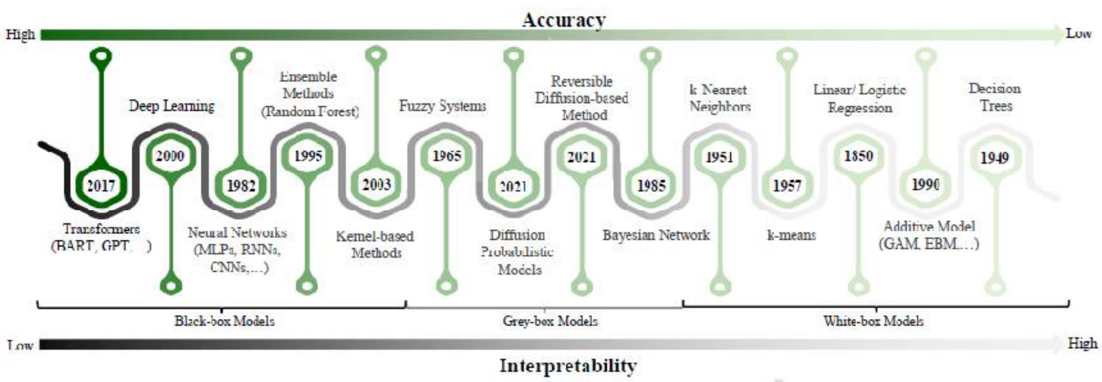
xAI Taxonomy



xAI.		2023-030

Relations among XAI concepts





xAI. Libraries

xAI.		2023-145

Some of popular XAI libraries

Name	Focus	Feature Importance	White-Box Models	Example-Based XAI	Visual XAI	Framework
AIX 360 [151]	General	LIME, SHAP	Decision Rules, Model Distillation	Prototypes, Contrastive Explanations	—	—
Alibi [152]	General	Anchors, Integrated Gradients, SHAP,	—	Contrastive Explanations, Counterfactuals	ALE	TensorFlow
Captum [153]	Neural Networks	DeepLift, Deconvolution, Integrated Gradients, SHAP, Guided Backpropagation, GradCam, Occlusion, PFI	—	—	—	PyTorch
DALEX [154]	General	LIME, SHAP, PFI	—	—	ALE, PDP	—
DiCE [155]	Counterfactuals	—	—	Counterfactuals	—	—
InterpretML [156]	General	LIME, SHAP, Morris Sensitivity Analysis	Explainable Boosting, Decision Tree, Decision Rules, Regression	—	PDP	—
PAIR Saliency [157]	Saliency Maps	Integrated Gradients, GradCam, Occlusion, Guided Backpropagation, Ranked Area Integrals, SmoothGrad	—	—	—	PyTorch, TensorFlow
Skater [158]	General	Layerwise Relevance Propagation, LIME, Integrated Gradients, Occlusion, PFI	Bayesian Rule List, Decision Tree	—	PDP	TensorFlow
Quantus [159]	Quantitative Evaluation	—	—	—	—	TensorFlow, PyTorch
ExplainerDashboard [160]	General	SHAP, PFI	Decision Tree	—	PDP	Scikit-learn
Ecco [161]	NLP	Integrated Gradients, Saliency, DeepLift, Guided Backprop	—	—	—	PyTorch
XAITK [162]	General	Saliency Maps	Decision Tree	Explanation by Example	—	—

xAI. 2023-155

Python frameworks supporting post-hoc attribution for XAI of DNNs

Framework	Back-end	Propagation Attribution	Propagation Rule-map	Other Attribution (Notable)	Documentation Tests
Zennit (ours)	PyTorch	Common LRP [10] Uncommon/Custom LRP Guided Backprop [37] Excitation Backprop [38]	Built-In Custom Canonization	SmoothGrad [35] Integrated Gradients [36] Occlusion [51]	Full Usage API Tutorials Fully Tested + CI
Captum [17]	PyTorch	LRP-ε [10] DeepLIFT(+Shap) [61, 62] Guided Backprop [37]	None	SmoothGrad [35] Integrated Gradients [36] Conductance [63, 64] GradientShap [62] KernelShap [62] GradCAM [65] Occlusion [51] LIME [66] Shapley Values [67, 68]	Full Usage API Tutorials Fully Tested + CI
TorchRay [30] (unmaintained)	PyTorch	Guided Backprop [37] Excitation Backprop [38]	None	GradCAM [65] Occlusion [51] LIME [66] RISE [69] Extremal Perturbation [30]	Joint Usage+API Examples Benchmarks
iNNvestigate [27]	Tensorflow/ Keras	Common LRP [10] PatternAttribution [70] DeepLIFT [61] Guided Backprop [37]	Built-In	SmoothGrad [35] Integrated Gradients [36]	Usage in Readme API Tutorials Fully Tested + CI
DeepExplain[71] (unmaintained)	Tensorflow/ Keras	LRP-ε [10] DeepLIFT [61]	None	Integrated Gradients [36] Occlusion [51] Shapley Values [67, 68]	Usage in Readme Examples Tests + CI

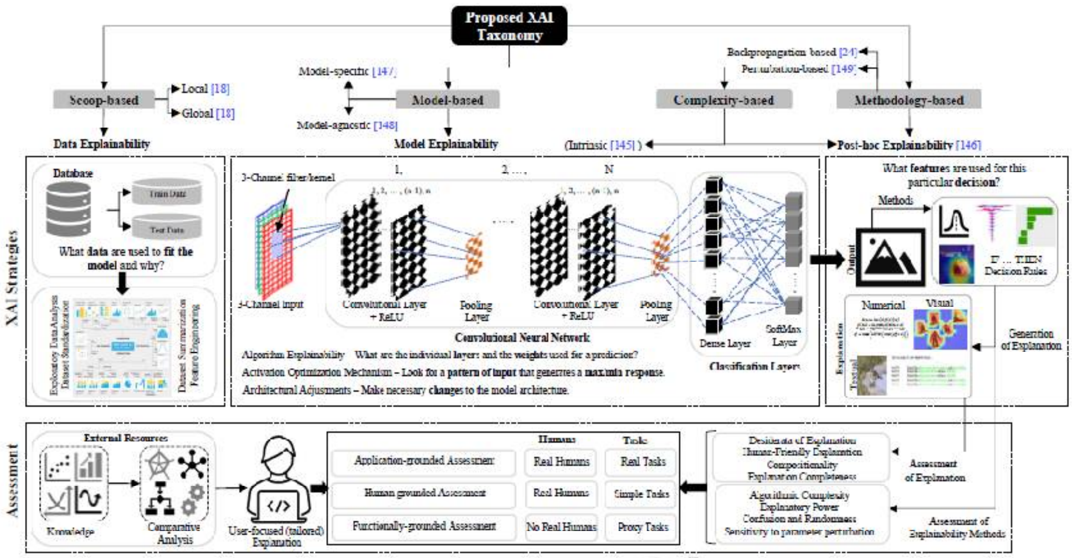
xAI. 2023-140

Explainability toolboxes

Toolbox	Publication	Code repository
Skater	Choudhary (2018)	https://github.com/oracle/Skater
InterpretML	Nori et al. (2019)	https://github.com/interpretml/interpret
iNNvestigate	Alber et al. (2019)	https://github.com/albermax/innvestigate
AI Fairness 360	Arya et al. (2019)	https://github.com/Trusted-AI/AIF360
explAIner	Spinner et al. (2020)	https://github.com/dbvis-ukon/explainer
FAT Forensics	Sokol et al. (2020)	https://github.com/fat-forensics/fat-forensics
Alibi	Klaise et al. (2021)	https://github.com/SeldonIO/alibi

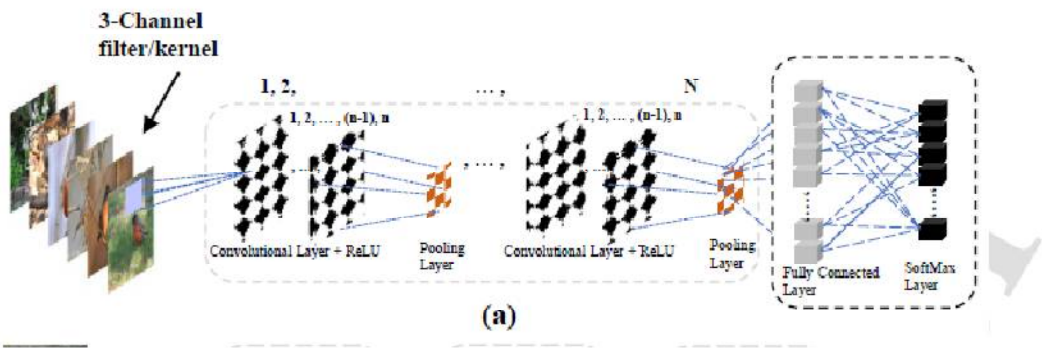
xAI. 2023-030

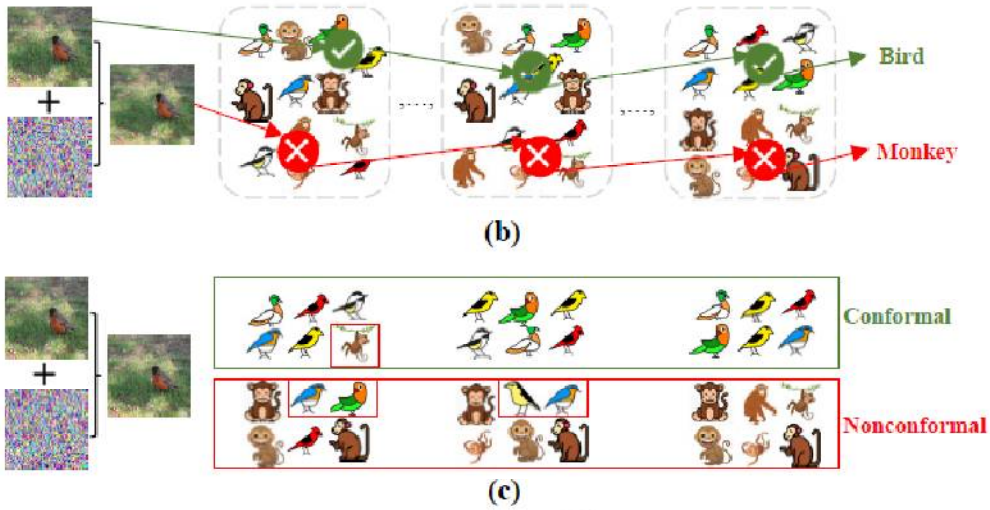
XAI methodology



xAI. 2023-030

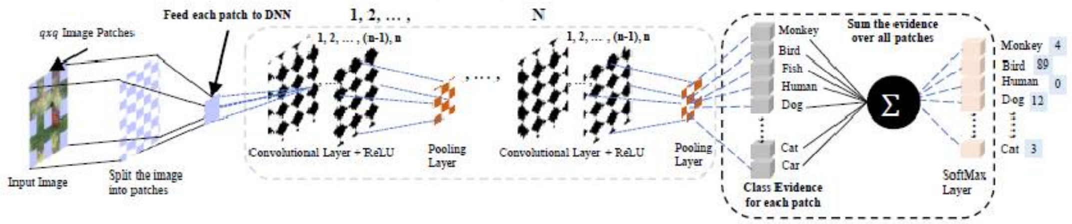
Deep NN





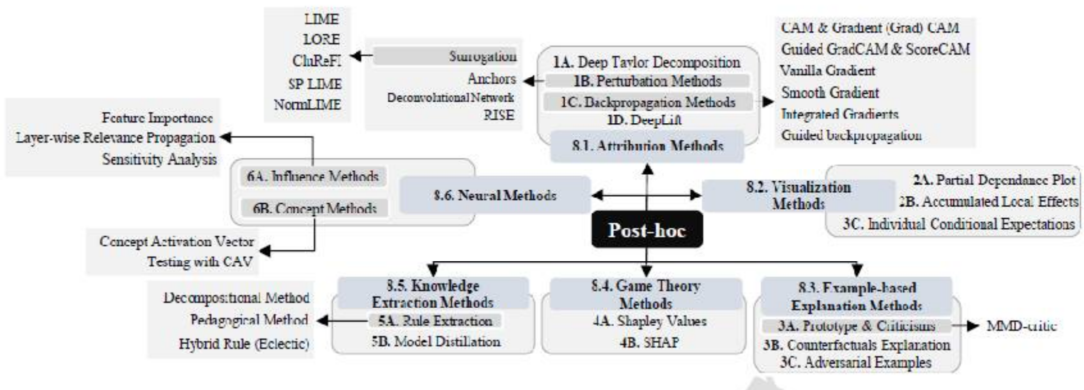
- ✓ (b) Output representations at each layer
- ✓ (c) Nearest neighbors at each layer.

BagNets



- Input is first split into q × q patches.
- Each patch is passed to DNN to extract the evidence score.
- Taken the sum of the class evidence scores overall patches to reach the final image classification decision

Post-hoc Explainability taxonomy

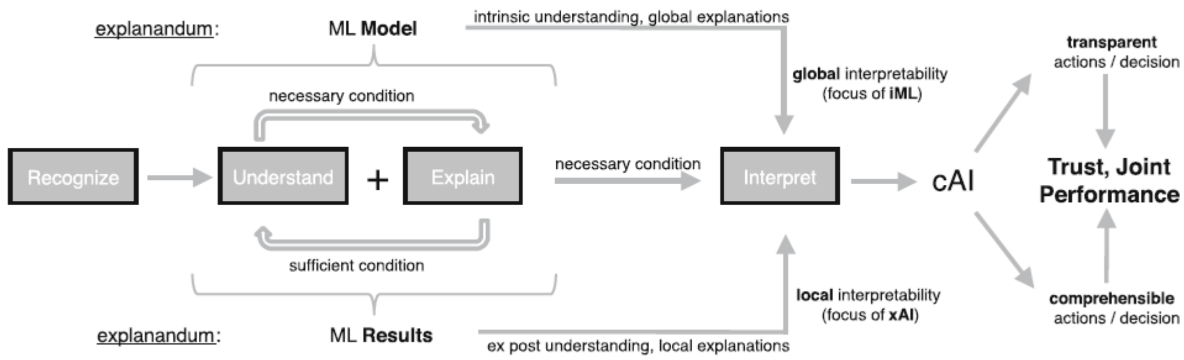


.

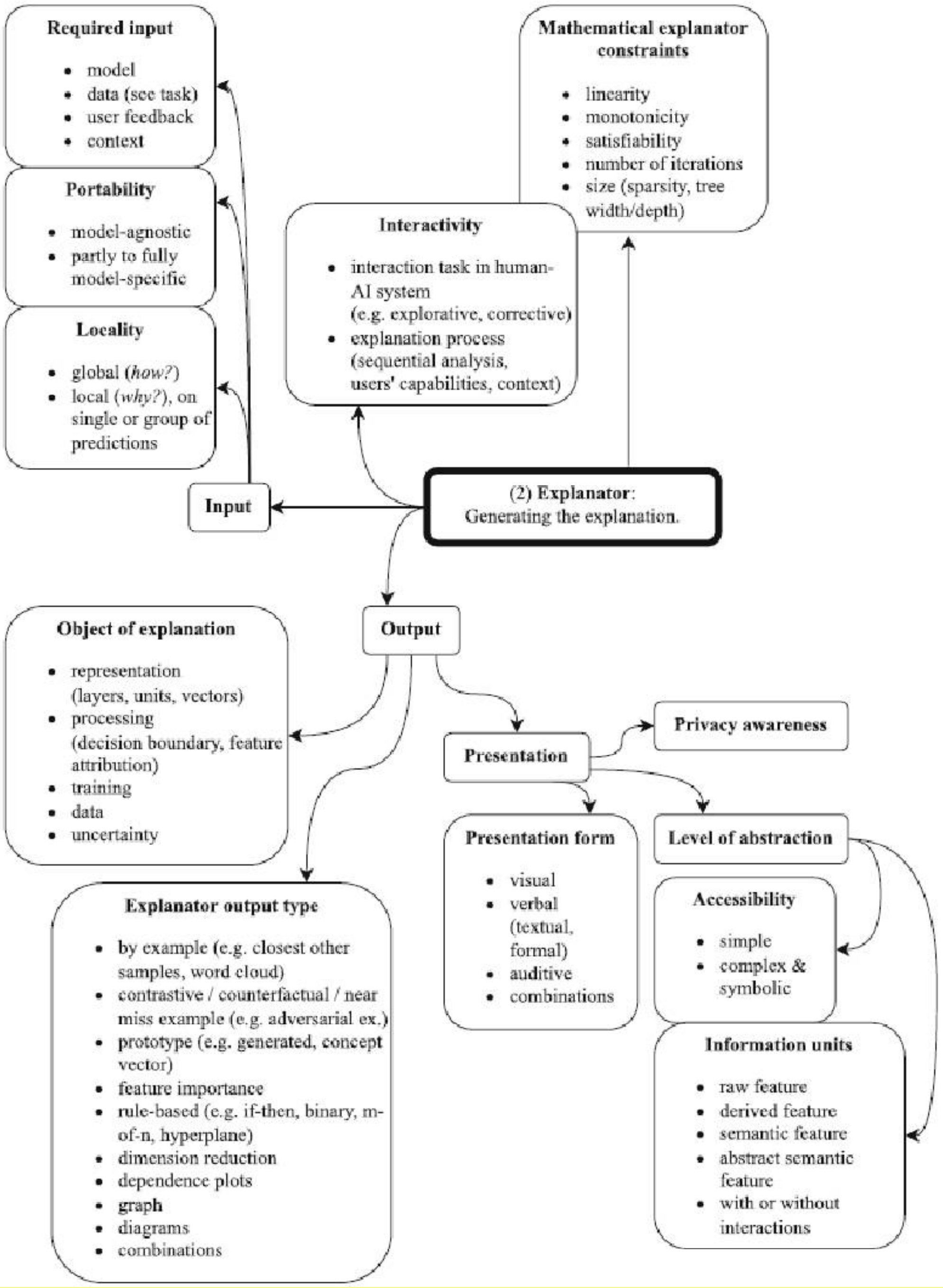
LIME	Locally Interpretable Model Agnostic Explainer
LIME.SP	Submodular Pick
LIME.RISE	Randomized Input Sampling to Provide Explanations,
LIME.CluRe	Cluster Representatives with LIME
LORE	Local Rule based Explanation
CAM	Class Activation Map
MMD	Maximum Mean Discrepancy
CAV	Concept Activation Vector

xAI.		2023-140

Framework for comprehensible AI

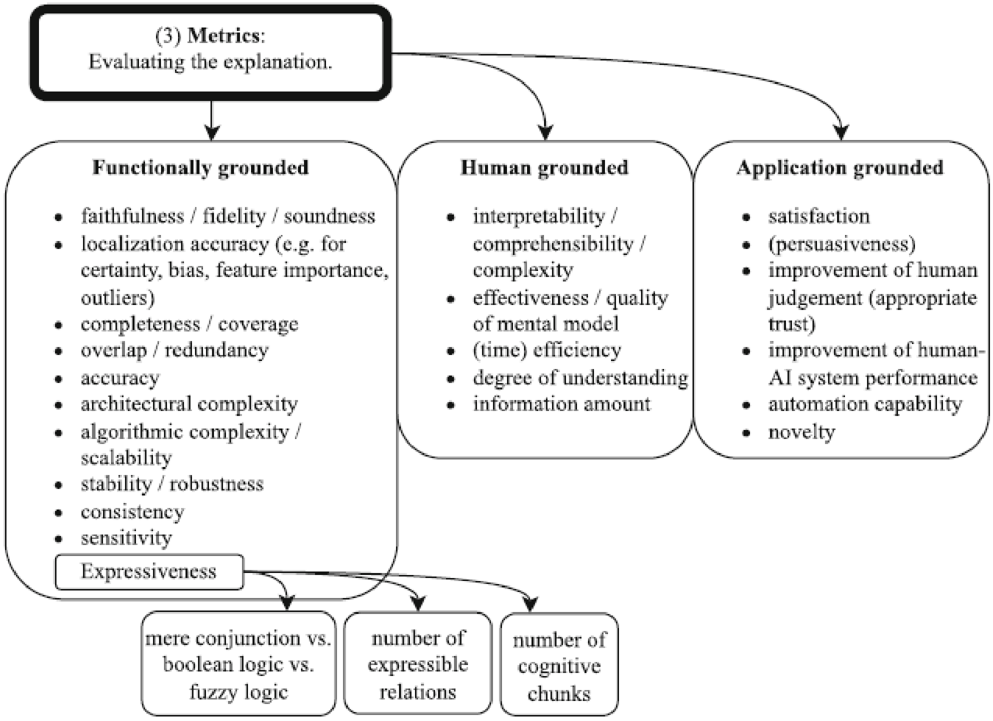


Explanator Taxonomy

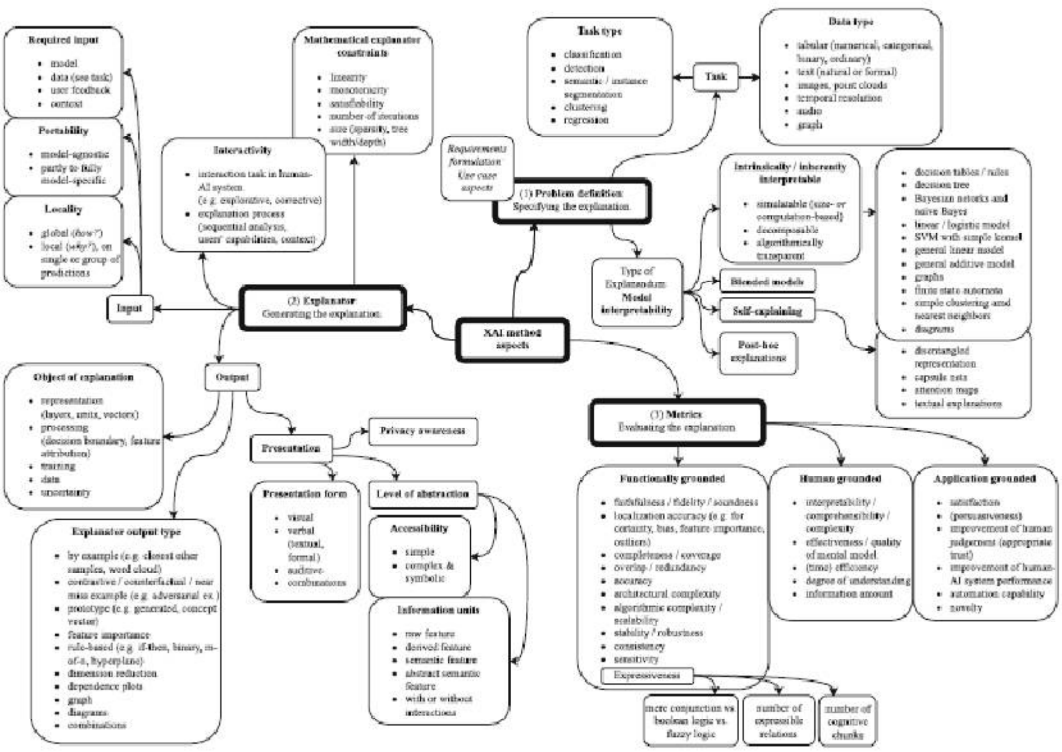


Required input
• model
• data (see task)
• user feedback
• context
Portability
• model-agnostic
• partly to fully model-specific
Locality
• global (how?)
• local (why?), on single or group of predictions
Mathematical explanator constraints
• linearity
• monotonicity
• satisfiability
• number of iterations
• size (sparsity, tree width/depth)
Interactivity
• interaction task in human-AI system (e.g. explorative, corrective)
• explanation process (sequential analysis, users' capabilities, context)
Input
(2) Explanator: Generating the explanation.
Output
Object of explanation
• representation (layers, units, vectors)
• processing (decision boundary, feature attribution)
• training
• data
• uncertainty
Presentation
Privacy awareness
Presentation form
• visual
• verbal (textual, formal)
• auditive
• combinations
Level of abstraction
Accessibility
• simple
• complex & symbolic
Information units
• raw feature
• derived feature
• semantic feature
• abstract semantic feature
• with or without interactions
Explanator output type
• by example (e.g. closest other samples, word cloud)
• contrastive / counterfactual / near miss example (e.g. adversarial ex.)
• prototype (e.g. generated, concept vector)
• feature importance
• rule-based (e.g. if-then, binary, m-of-n, hyperplane)
• dimension reduction
• dependence plots
• graph
• diagrams
• combinations

XAI metrics



Complete taxonomy of xAI



xAI-Probes

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Self-explaining and blended models								
–	Hendricks et al. (2016)	cls		s		p	sym/vis	rules/fi
–	Kim et al. (2018b)	any		s		p	sym/vis	rules/fi
ProtoPNet	Chen et al. (2019a)	cls,img		s		p/r	vis	proto/fi
Capsule Nets	Sabour et al. (2017)	cls		s		r	sym	fi
Semantic Bottlenecks, ReNN, Concept Whitening	Losch et al. (2019), Wang (2018), Chen et al. (2020)	any		s		r	sym	fi
Logic Tensor Nets	Donadello et al. (2017)	any		b	✓	p/r	sym	rule
FoldingNet	Yang et al. (2017)	any,pcl		b		p	vis	fi/red
Neuralized clustering	Kauffmann et al. (2019)	any		b		p	vis	fi
Black-box heatmapping								
LIME, SHAP	Ribeiro et al. (2016), Lundberg and Lee (2017)	cls	✓	p		p	vis	fi/con
RISE	Petsiuk et al. (2018)	cls,img	✓	p		p	vis	fi
D-RISE	Petsiuk et al. (2021)	det,img	✓	p		p	vis	fi
CEM	Dhurandhar et al. (2018)	cls,img	✓	p		p	vis	fi/con
White-box heatmapping								
Sensitivity analysis	Baehrens et al. (2010)	cls		p		p	vis	fi
Deconvnet, (Guided) Backprop.	Zeiler and Fergus (2014), Simonyan et al. (2014), Springenberg et al. (2015)	img		p		p	vis	fi
CAM, Grad-CAM	Zhou et al. (2016), Selvaraju et al. (2017)	cls,img		p		p	vis	fi
SIDU	Muddamsetty et al. (2021)	cls,img		p		p	vis	fi
Concept-wise Grad-CAM	Zhou et al. (2018)	cls,img		p		p/r	vis	fi
SIDU	Muddamsetty et al. (2021)	cls,img		p		p	vis	fi
LRP	Bach et al. (2015)	cls		p		p	vis	fi

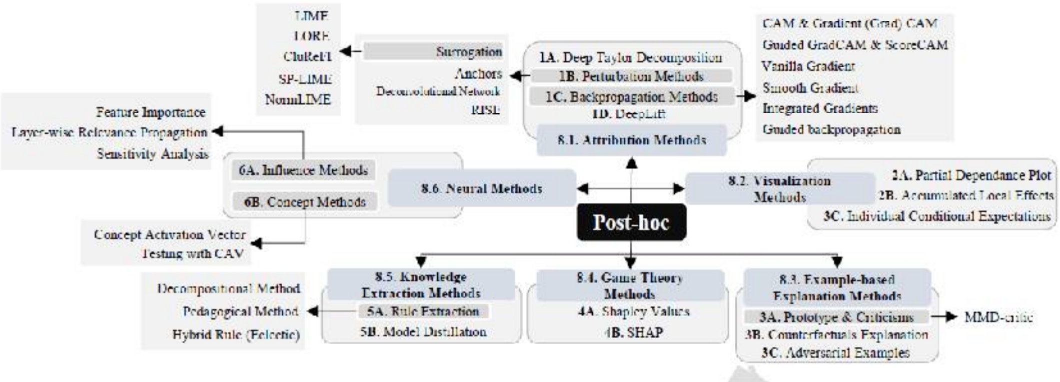
Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Pattern attribution	Kindermans et al. (2018)	cls		p		p	vis	fi
–	Fong and Vedaldi (2017)	cls		p		p	vis	fi
SmoothGrad, Integrated Gradients	Smilkov et al. (2017), Sundararajan et al. (2017)	cls		p		p	vis	fi
Integrated Hessians	Janizek et al. (2020)	cls		p		p	vis	fi
Global representation analysis								
Feature Visualization	Olah et al. (2017)	img		p	✓	r	vis	proto
NetDissect	Bau et al. (2017)	img		p	✓	r	vis	proto/fi
Net2Vec	Fong and Vedaldi (2018)	img		p	(✓)	r	vis	fi
TCAV	Kim et al. (2018a)	any		p	✓	r	vis	fi
ACE	Ghorbani et al. (2019)	any		p	✓	r	vis	fi
–	Yeh et al. (2020)	any		p	✓	r	vis	proto
IIN	Esser et al. (2020)	any		p	(✓)	r	vis/sym	fi
Explanatory Graph	Zhang et al. (2018)	img		p	(✓)	p/r	vis	graph
Dependency plots								
PDP	Friedman (2001)	any	✓	p		p	vis	plt
ICE	Goldstein et al. (2015)	any	✓	p	✓	p	vis	plt
Rule extraction								
TREPAN, C4.5, Concept Tree	Craven and Shavlik (1995), Quinlan (1993), Renard et al. (2019)	cls	✓	p	✓	p	sym	tree
VIA	Thrun (1995)	cls	✓	p	✓	p	sym	rules
DeepRED	Zilke et al. (2016)	cls		p	✓	p	sym	rules
LIME-Aleph	Rabold et al. (2018)	cls	✓	p		p	sym	rules
CA-ILP	Rabold et al. (2020)	cls		p	✓	p	sym	rules
NBDT	Wan et al. (2020)	cls		p	✓	p	sym	tree

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Interactivity								
CAIPI	Teso and Kersting (2019)	cls,img	✓	p		r	vis	fi/con
EluciDebug	Kulesza et al. (2010)	cls	✓	p		r	vis	fi,plt
Crayons	Fails and Olsen Jr (2003)	cls,img	✓	t		p	vis	plt
LearnWithME	Schmid and Finzel (2020)	cls	✓	t	✓	p, r	sym	rules
Multi-modal phrase-critic model	Hendricks et al. (2018)	cls,img		p	✓	p	vis,sym	plt,rules
Inspection of the training								
–	Shwartz-Ziv and Tishby (2017)	any		p	✓	t	vis	dist
Influence functions	Koh and Liang (2017)	cls		p	✓	t	vis	fi/dist
Data analysis methods								
t-SNE, PCA	van der Maaten and Hinton (2008), Jolliffe (2002)	any	✓	p	✓	d	vis	red
k-means, spectral clustering	Hartigan and Wong (1979), von Luxburg (2007)	any	✓	p	✓	d	vis	proto

Abbreviations by column: *image data*=img, *point cloud data*=pcl; *Trans.*=transparency, *post-hoc*=p, *transparent*=t, *self-explaining*=s, *blended*=b; *processing*=p, *representation*=r, *development during training*=t *data*=d; *visual*=vis, *symbolic*=sym, *plot*=plt; *feature importance*=fi, *contrastive*=con, *prototypical*=proto, *decision tree*=tree, *distribution*=dist

xAI. | 2023-030

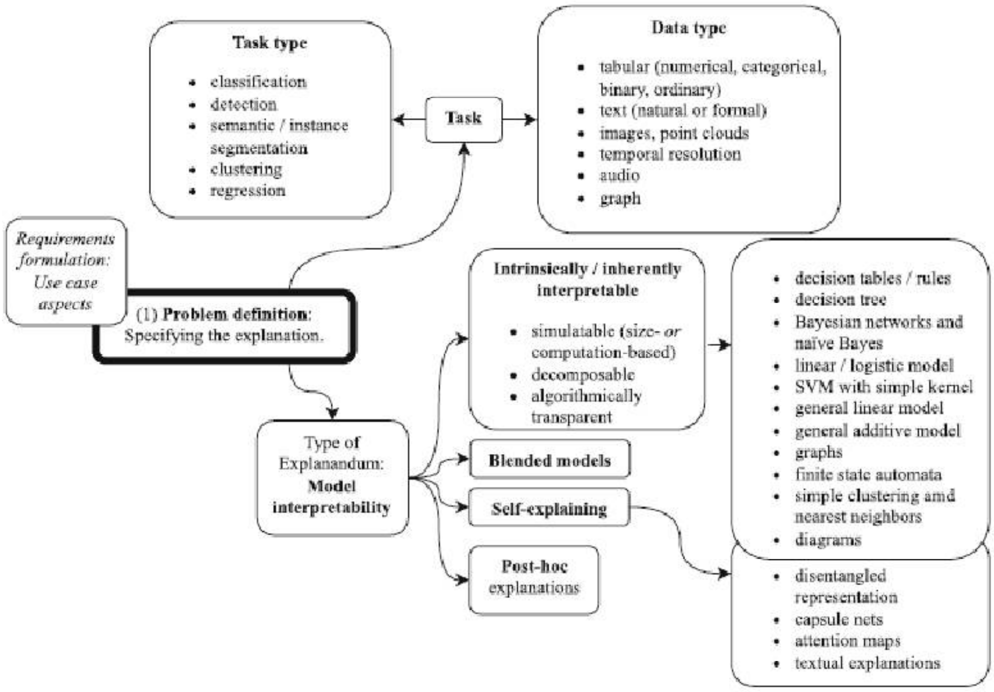
Post-hoc Explainability



xAI. | 2023-030

A comprehensive overview of attribution-based XAI methods, highlighting advantages and disadvantages.

Method	Ref.	Advantages	Disadvantages	Concept
DTD	[240]	Training free method, may apply directly to any NNs.	i) Inconsistent in providing a unique solution, and slow computations [245]; ii) Partial explanation as higher order derivatives terms are set to zeros.	SA methods
LIME	[29]	i) Suitable to a very large number of explanatory variables, sparse explainer; ii) Same local interpretable model could be replaced [149]; iii) Selective and possibly contrastive explanations; iv) Provides local fidelity; v) Makes no assumptions about the model.	i) Incapable of explaining models with non-linear decision boundaries; ii) Incapable of explaining surrounding observations [149]; iii) Unsolved problem with tabular data.	Model agnostic local surrogate
LORE	[246]	i) Provide a counterfactual suggestion with the explanation; ii) Utilise a genetic algorithm that takes advantage of the black-box to generate examples; iii) Parameter-free method.	i) Based on assumption; ii) Cannot provide a global explanation; iii) Works for tabular data.	Local explanation
CluReFi	[247]	Provides local explanation to a cluster.	Representative of each cluster presents the explanation of important features.	Local explanation
SP-LIME	[29]	To check the entire model by extracting some data points. Aggregate the local models to form a global interpretation.	Less beneficial for high-level comprehension.	Model agnostic global surrogate
NormLIME	[227]	Provides finer-grained interpretation in a multi-class setting and add proper normalization to reduce the computation.	Aggregate many explanations for the class-specific explanation.	Local explanation
Anchors	[231]	i) Less computation than SHAP; ii) Better generalizability than LIME [227].	i) Requires discretization, highly configurable, and impactful setup; ii) Coverage drastically decreases with an increase in the number of feature predicates.	Perturbation-based model agnostic RL
DeconvNet	[238]	i) Highlights fine-grained details; ii) Dense feature representation with multi-layer.	Artifacts in the visualization [31]; ii) Training is difficult due to the large output space.	Pixel-space gradient visualization
RISE	[229]	i) Any architecture can be generalized; ii) Proposes causal metrics.	i) Inconsistent due to random mask; ii) Slow computation.	Pixel saliency
CAM	[235]	i) Identifies discriminative areas in an image classification task; ii) Fast and accurate.	i) Modify the network architecture that lends to complex model [31]; ii) Applicable to a specific type of CNN.	Regularization
Grad CAM	[31]	i) Applies to a broad range of CNN model families; ii) Robust to adversarial perturbations in an image classification task; iii) Help to achieve the model generalization by removing biases.	i) Lacks the ability to highlight fine-grained details; ii) Individual interpretations are difficult to aggregate for global knowledge.	Regularization
Guided Backpropagation	[240]	i) Highlights the fine-grained details and less noisy explanation [31]; ii) Provides more interpretable results than DeepLift.	i) Captures pixels detected by neurons, not the ones that suppress neurons [31]; ii) Less class-sensitive than the vanilla gradient.	Pixel-space Gradient Visualization
Guided Grad-CAM	[31]	i) Removes negative gradients and understand the model's decision; ii) Provides class descriptive and high-resolution maps.	i) Distinguishes an object of the same class; ii) Does not consider the entire class region.	Guided Back-propagation + Grad-CAM
ScoreCAM	[30]	i) Solves the dependency's problem on the gradients; ii) Achieves better visualization and fair interpretation.	i) Localization results are poor and lead to non-interpretability; ii) Smoothing generates inconsistent explanations.	CAM
Vanilla Gradient	[239]	i) Simple to implement based on backpropagation; ii) Pixel-wise features are important.	i) Makes undesirable changes with data pre processing [249]; ii) Vulnerable to adversarial attacks [250]; iii) Decision-making process is unknown.	Backpropagation interpretation
SmGrad	[233]	i) Denoising impact on the sensitivity map is achieved by training with noisy data; ii) Generates images with multiple levels of noise.	i) More effective with Large areas of the class object. ii) Degeneralizes to different networks.	Regularization [251]
Integrated Gradients (IG)	[252]	i) Very suitable for neural networks; ii) Optimizes the heatmap for faithful explanations.	i) Does not meet the Shapley values' axiom; ii) Frail mechanism to identify specific features and inconsistent to produce the explanation.	Shapley value
DeepLift	[234]	i) Gradient-free [227]; ii) Achieves the goal of completeness.	i) Depends on a reference point or baseline; ii) Produces inconsistent results due to redefining gradient.	Feature importance



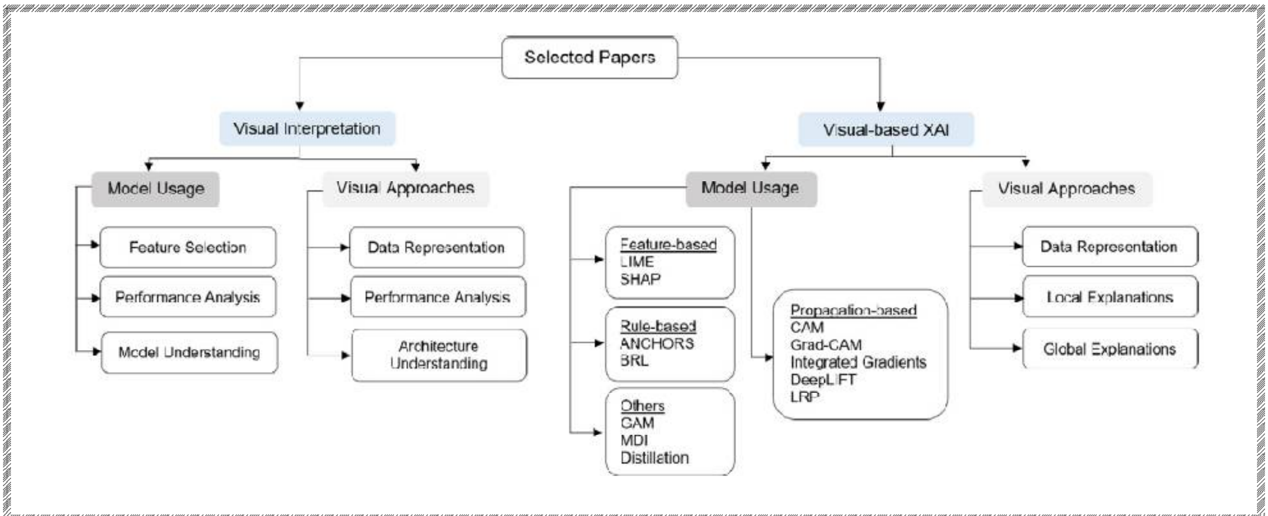
xAI.		2022-183

Classification of some of popular XAI probes

XAI method	Explanation level		Implementation level		Model dependency	
	Global	Local	Intrinsic	Post hoc	Agnostic	Specific
ANCHORS [40]		✔		✔	✔	
LIME [28]	✔	✔		✔	✔	
SHAP [35]		✔		✔	✔	
LRP [30]	✔	✔		✔	✔	
Grad-CAM [29]		✔		✔	✔	
Saliency Maps [39]		✔		✔	✔	
Integrated Gradients [38]		✔		✔	✔	
DeepLIFT [36]		✔		✔	✔	
Bayesian Rule Lists [32]	✔		✔			✔
Distillation [34]	✔			✔	✔	
GAM [33]	✔		✔			✔
Mean Decrease Impurity [37]	✔	✔	✔			✔
CAM [41]		✔		✔	✔	

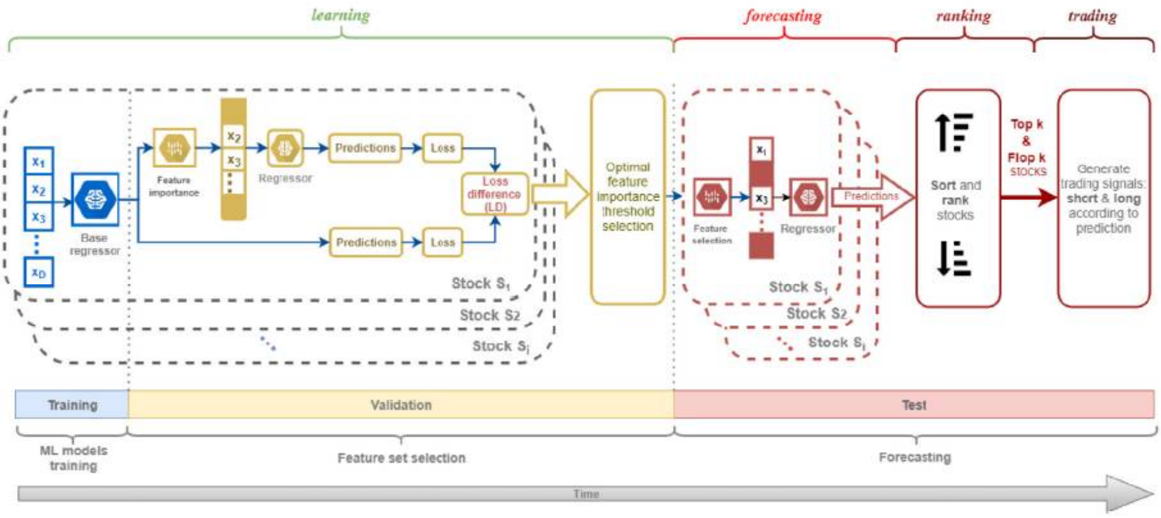
xAI.		2022-183

XAI probes



xAI.	XAI StatArb package	2022-015 2022-044

Block diagram of the StatArb XAI trading strategy



Feature selection strategies for three stocks:
GOOGLE, IBM, and INTC

Stock	Feature	Feature importance
GOOGL	LR_2	-0.1077
GOOGL	LR_{63}	-0.3906
GOOGL	LR_5	-0.5298
GOOGL	LR_3	-1.1571
IBM	LR_1	0.1060
INTC	LR_5	-0.2911
INTC	LR_{252}	-0.3211
INTC	LR_3	-0.3420
INTC	LR_{63}	-0.7658
INTC	LR_{126}	-1.1547

(a) Feature importance

	PI best		PI worst		PI running	
Stock	Feature	Feature importance	Feature	Feature importance	Feature	Feature importance
GOOGL	N/A		LR_3	-1.1571	LR_{63}	-0.3906
IBM	N/A		N/A		N/A	
INTC	LR_5	-0.2911	LR_{126}	-1.1547	LR_5	-0.2911

(b) Selected features

xAI.		2022-103

Perturbation

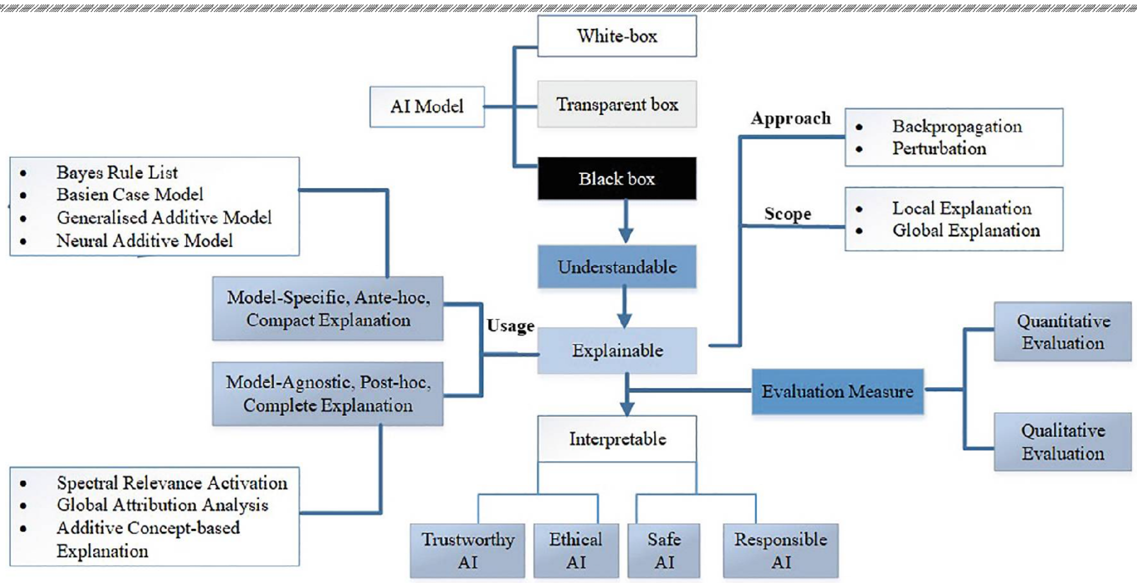
	Method	References
Gradients (sensitivity)	N/A (gradient-based)	Baehrens et al., 2009
	Saliency maps	Simonyan et al., 2013
	Class activation mapping (CAM)	Zhou et al., 2016
	Gradient-weighted CAM (Grad-CAM)	Selvaraju et al., 2017
	Guided Grad-CAM	Selvaraju et al., 2016
	3D CAM	Yang et al., 2018
	3D Grad-CAM	Yang et al., 2018
	Respond-CAM	Zhao et al., 2018
	Multiscale CAM	Hu et al., 2020
	SmoothGrad (SG)	Smilkov et al., 2017
	Correlation maps	Schirrmeister et al., 2017
	Testing with concept activation vectors (TCAV)	Kim et al., 2018
	Automated concept-based explanation (ACE)	Ghorbani et al., 2019a,b
Signal	Guided backpropagation (GBP)	Springenberg et al., 2014
	DeConvNet (occlusion maps)	Zeiler and Fergus, 2014
	Inversion-based	Mahendran and Vedaldi, 2015
	Inversion-based	Dosovitskiy and Brox, 2016
	PatternNet	Kindermans et al., 2017
	PatternAttribution	Kindermans et al., 2017
Model agnostic	Local interpretable model-agnostic explanations (LIME)	Ribeiro et al., 2016
	Submodular pick LIME (SP LIME)	Ribeiro et al., 2016
	anchor-LIME (aLIME)	Tulio Ribeiro et al., 2016
	Model agnostic globally interpretable explanations	Puri et al., 2017
	SHapley additive exPlanations (SHAP)	Lundberg and Lee, 2017

Decomposition (redistribution)

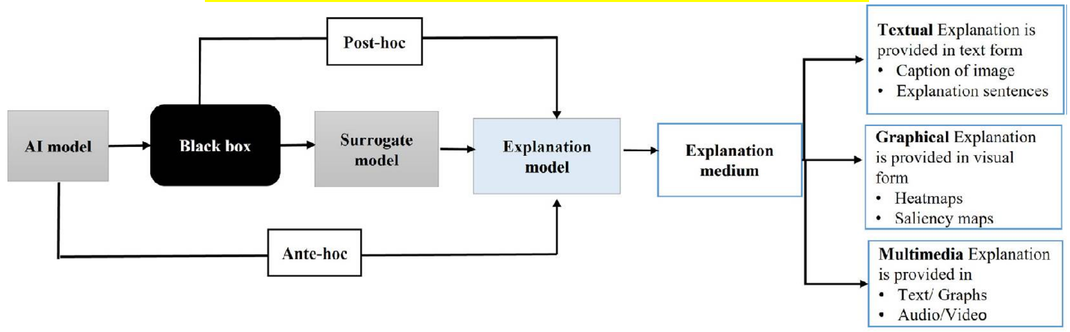
Layer-wise relevance propagation (LRP)	Bach et al., 2015
Deep Taylor decomposition	Montavon et al., 2017
Deep learning important FeaTures (DeepLIFT)	Shrikumar et al., 2017
Integrated gradients (IG)	Sundararajan et al., 2017
Gradient $\times$ input	Shrikumar et al., 2017
Prediction difference analysis (PDA)	Zintgraf et al., 2017
Graph LRP	Chereda et al., 2021

xAI.		2022-045

XAI methods terminology map



Workflow of Post-hoc and Ante-hoc Global XAI methods

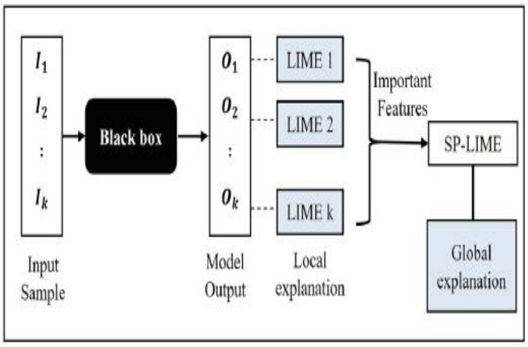


Model usage (type)	Year	Methods	Data type
Ante-hoc (Model-Specific)	2014	BCM [75]	Any
	2015	GAM [76]	Tabular
	2015	BRL [86]	Tabular
	2020	NAM [70]	Image

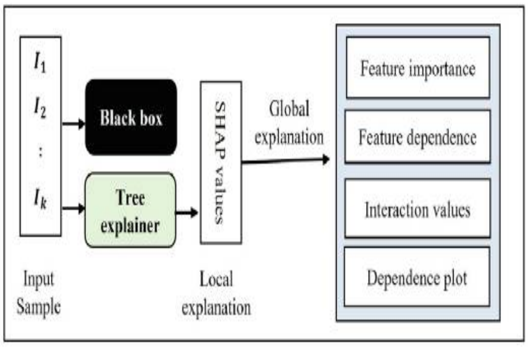
Methodologies	Explanation medium	Frameworks	XAI evaluation
Perturbation-based	Multimedia	Python (PYMC)	Qualitative
Perturbation-based	Graphics (heatmaps)	R (PyGAM)	Qualitative
Rule-based	Textual	Python	Quantitative
Cluster-based	Graphics (heatmaps)	Pytorch	Quantitative

Model usage (type)	Year	Methods	Data type	Methodologies	Explanation medium	Frameworks	XAI evaluation
Post-hoc (Model-Agnostic)	2016	SP-LIME[40]	Any	Perturbation-based	Graphics	Python/R	Qualitative
	2015	LRP [86]	Image	Gradient-based	Graphics (heatmaps)	Caffe	Quantitative
	2017	SHAP [52]	Any	Perturbation-based	Multimedia	Python (XGBoost)	Quantitative
	2019	SpRAy [53]	Image	Gradient-based	Graphics	Caffe	Quantitative
	2019	GAA [72]	Image	Perturbation-based	Multimedia	Multi- dimensional	Quantitative
	2019	ACE [94]	Image	Concept-based	Graphics	TensorFlow	Qualitative

Global explanation from Local interpretation SP-LIME and SHAP methods

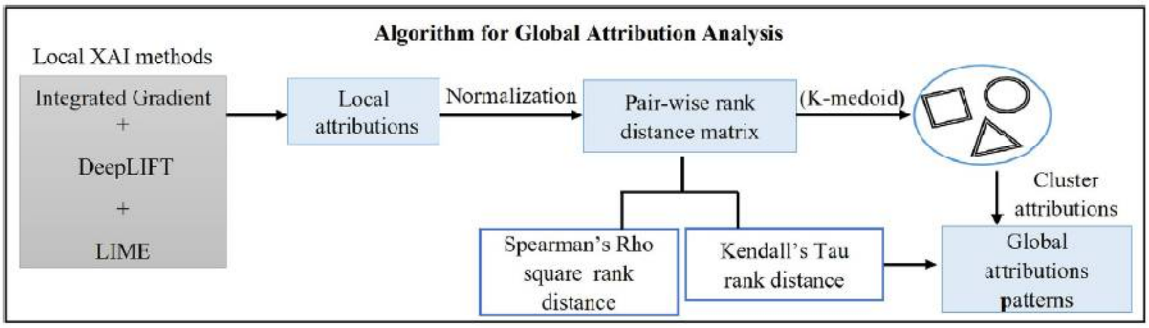


(a) SP-LIME (global explanation) from LIME (local explanation)

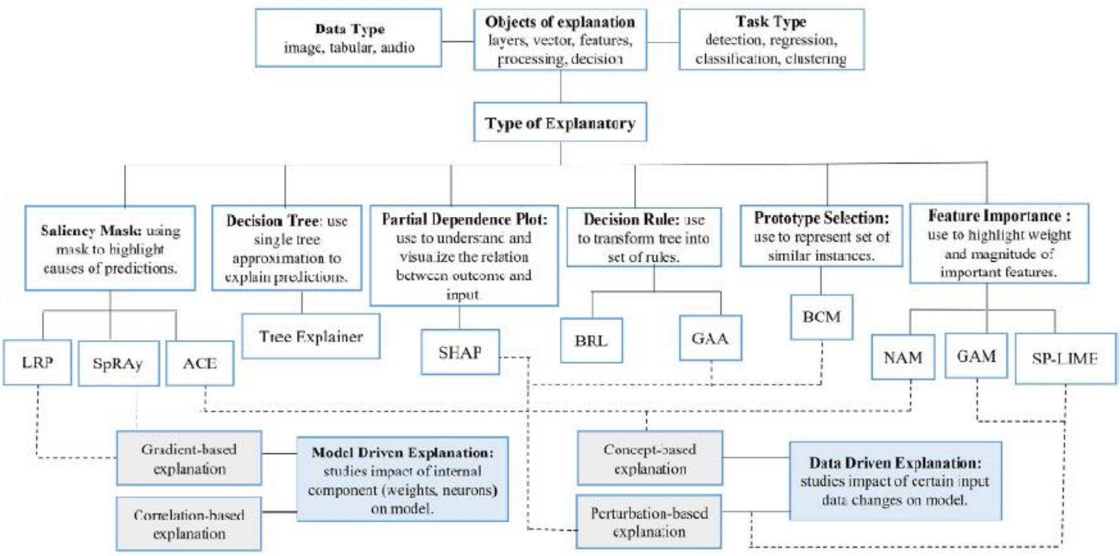


(b) Combining different local explanations and from SHAP values for Global explanation

Algorithm for Global Attribution Analysis for global interpretation



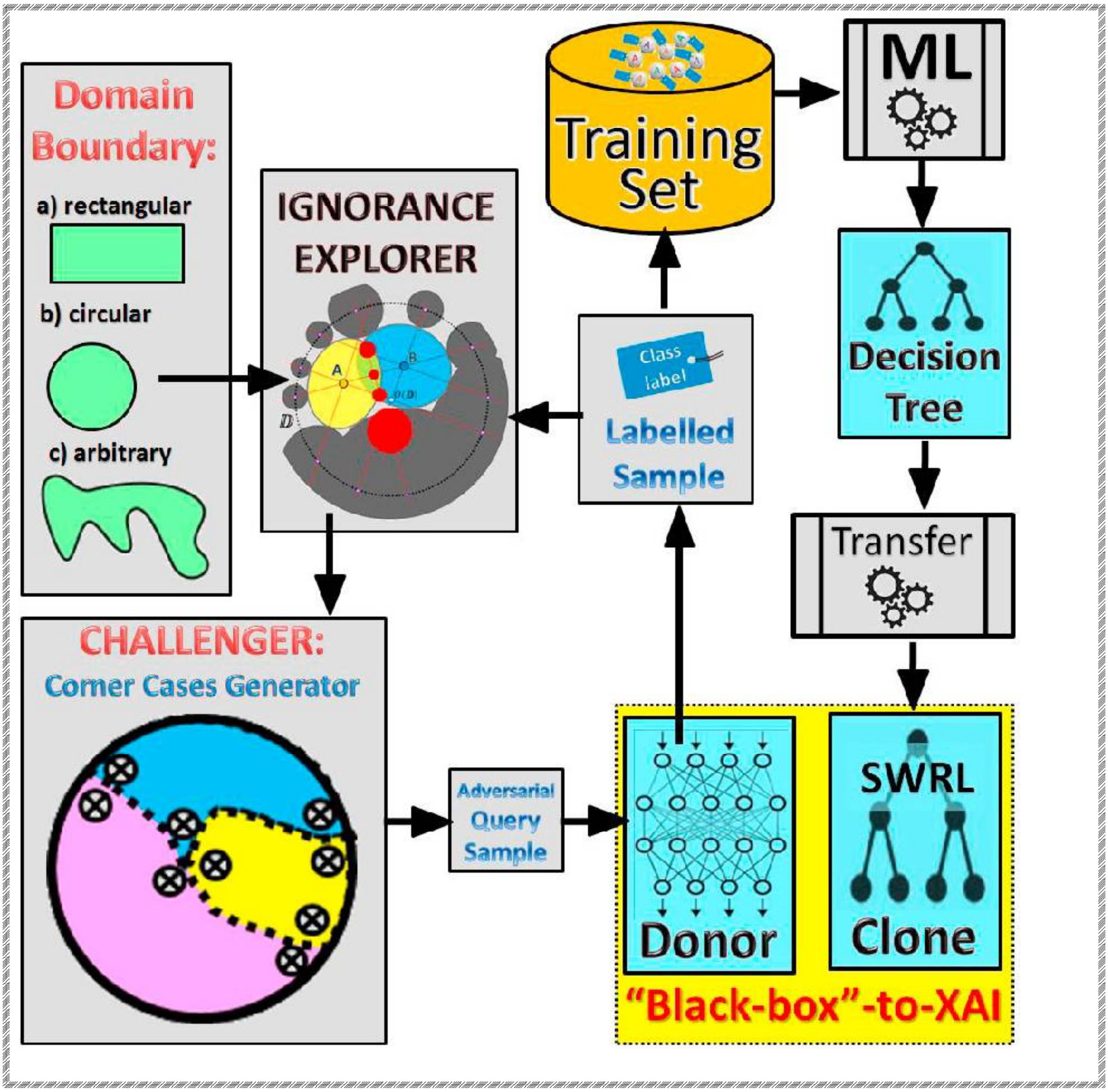
Explanatory Taxonomy of
Data and Model Driven Global Explainable methods



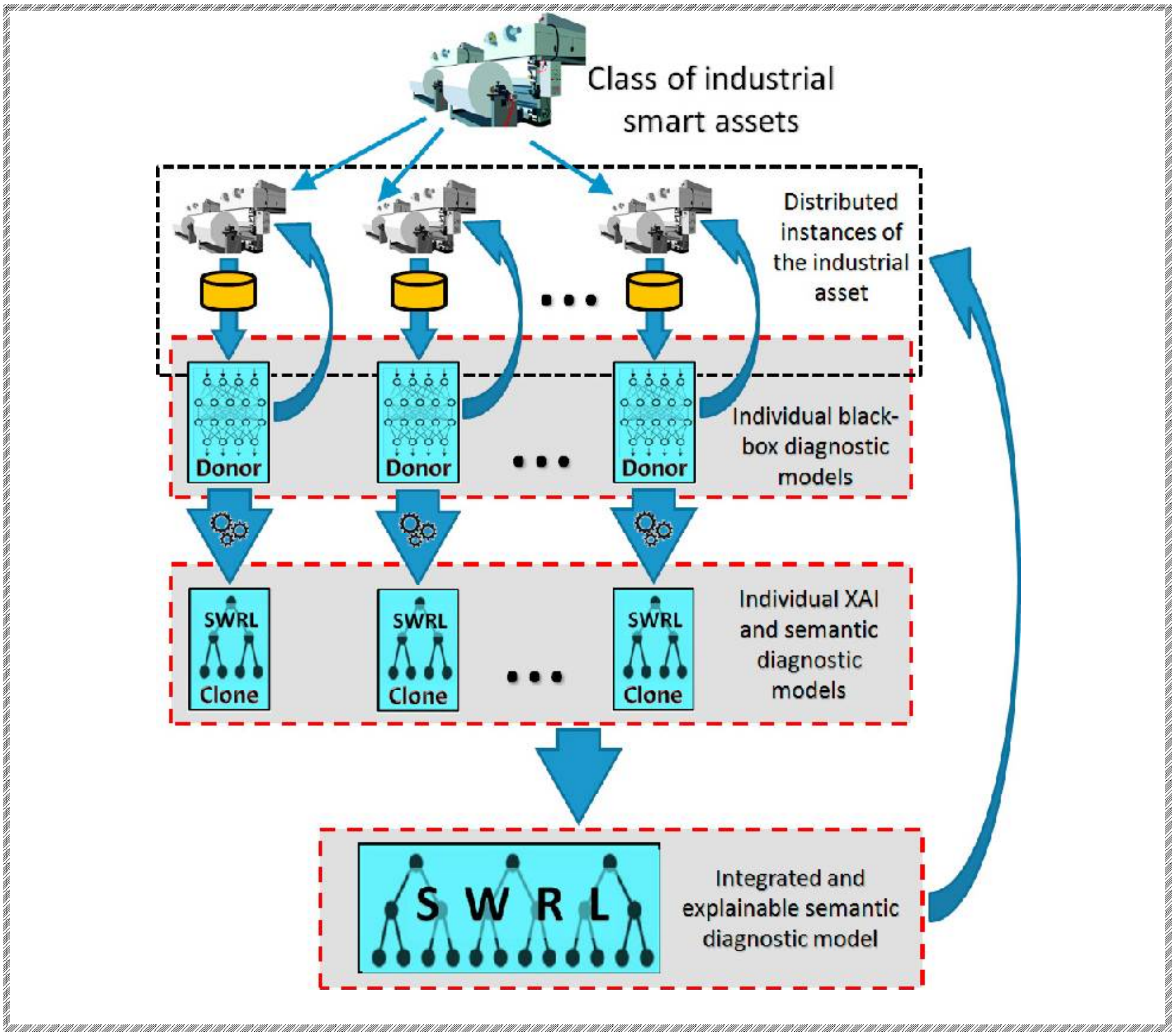
Explanators	Merits	Demerits
Saliency Map	• Highlight important pixels. • Faster computation.	• Only qualitative evaluation is available. • Insensitive to model and data.
Decision Tree	• Easy to explain. • Need less effort for data preparation.	• Fail to deal with linear relationships. • Difficult and expensive to interpret deeper tree.
Partial Dependence Plot	• Easy to understand and interpret. • Easy to implement.	• Deal with maximum two features. • Hidden Heterogeneous effect.
Decision Rule	• Select only the relevant features. • Cost efficient.	• Difficult and tedious to list all the rules. • Fail to describe linear relationship.
Prototype Selection	• Easy detection of missing functionality. • Detect error at early stage.	• Expensive. • Higher number of features or clusters.
Feature Importance	• Fast interpretation. • Highly compressed and insight model globally.	• Expensive. • Time consuming.

xAI.	Semantic Web Rule Language (SWRL) Semantic Representation of Deep Learning Models (SenRepDLM)	2022-104

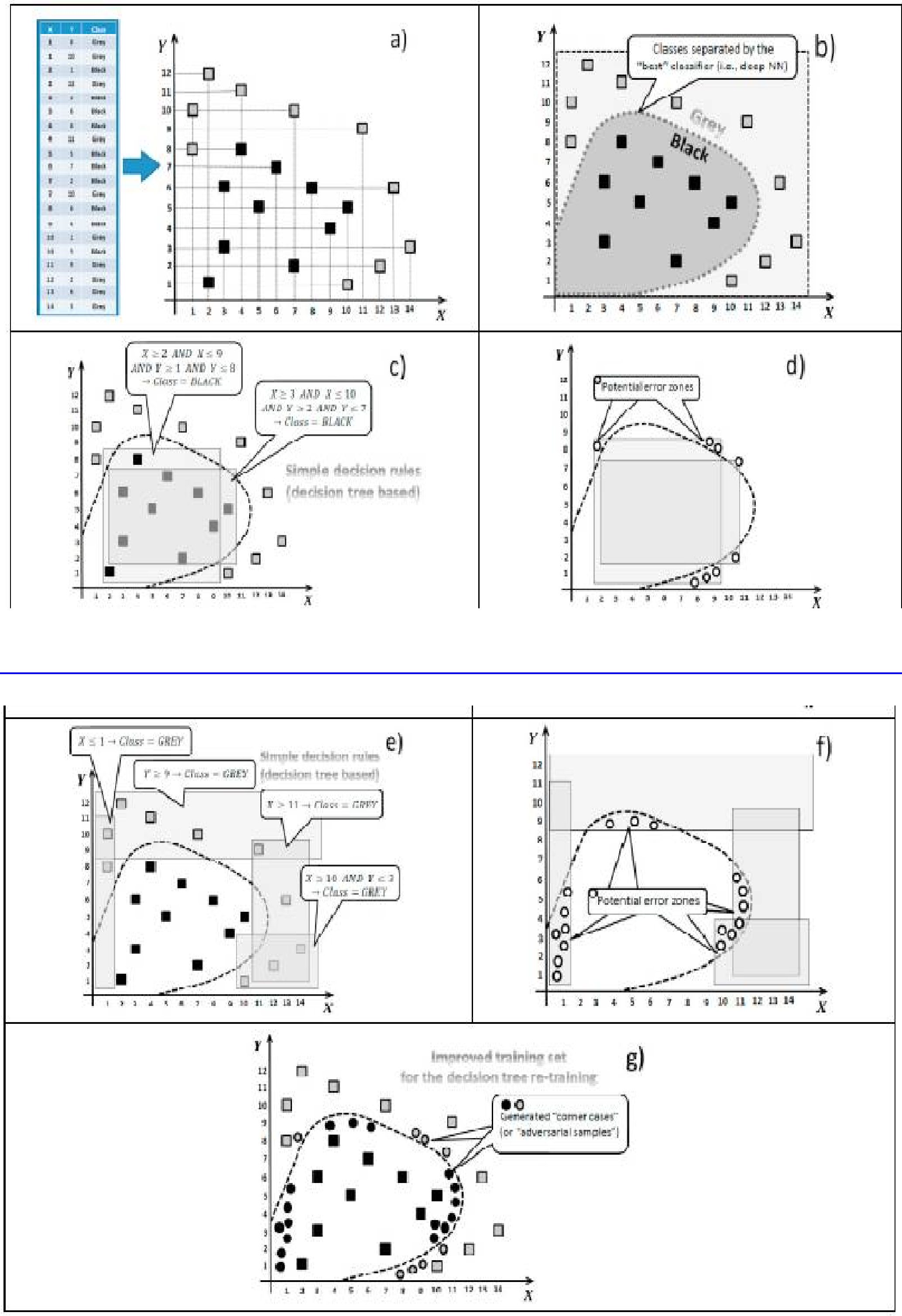
Generic schema of "cloning" black-box classification models
to the explainable form of SWRL rules



xAI.		2022-104
Black-box – XAI (SWRL) transformation and integration Use case scenario (predictive maintenance of smart industrial assets)		



xAI.		2022-104
Neural network vs. decision tree example		

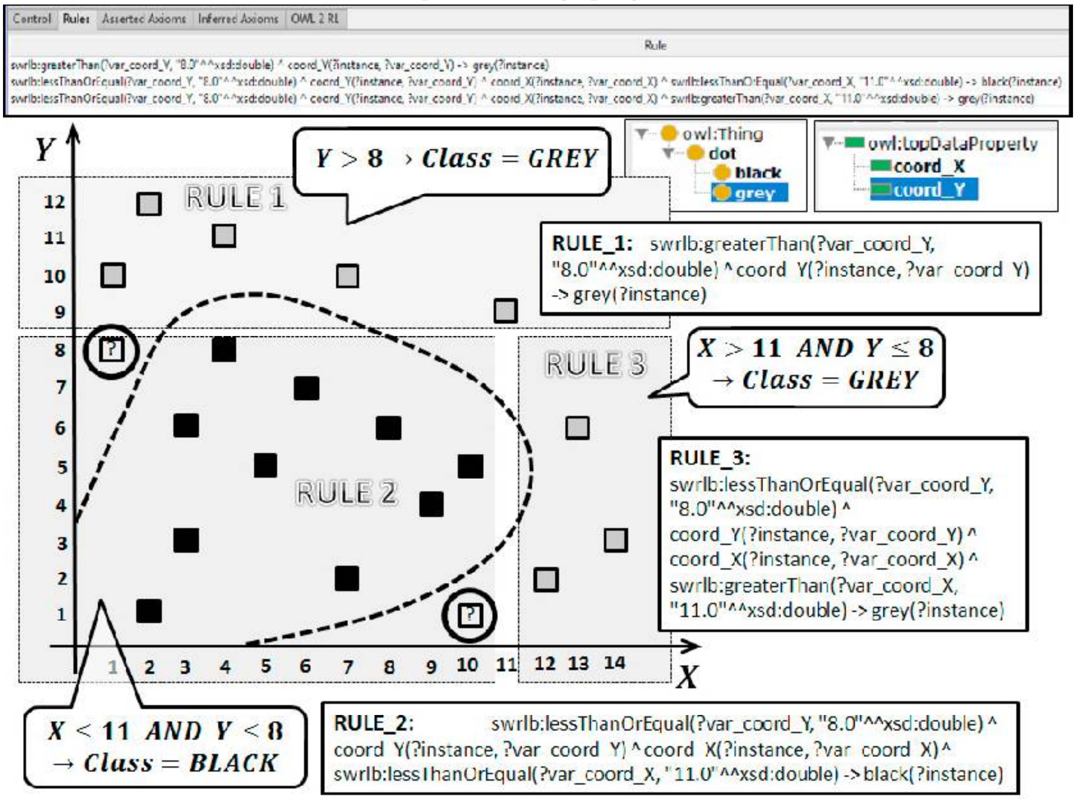


(a) 20 Training samples;
(b) Decision boundary produced by neural network;
(c) Two rules for the class "black" produced by the trained decision tree;
(d) Potential error zones for the decision tree (regarding class "Black") where the trained

neural network performs better;
(e) Two rules for the class “Grey” produced by the trained decision tree;.
(f) Potential error zones for the decision tree (regarding class “Grey”) where the trained neural network performs better;
(g) Generated samples (“corner cases” or “adversarial samples”) within the discovered potential error zones that could be used to re-train the decision tree aiming better classification accuracy and robustness.

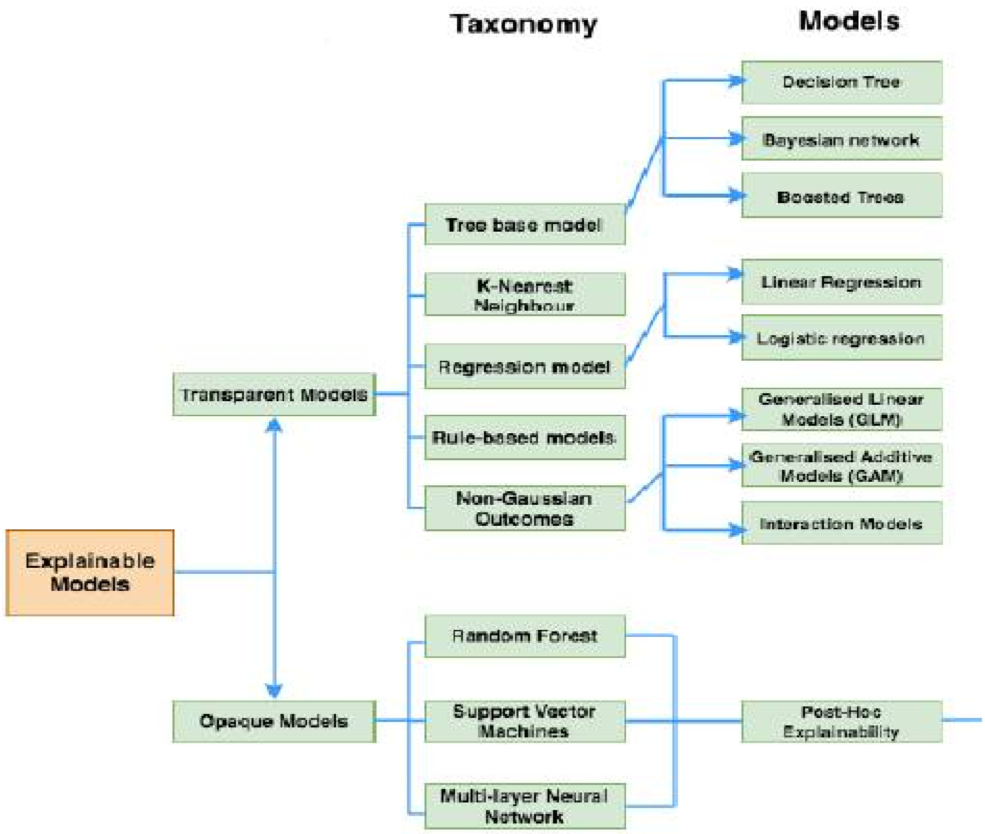
xAI.		2022-104

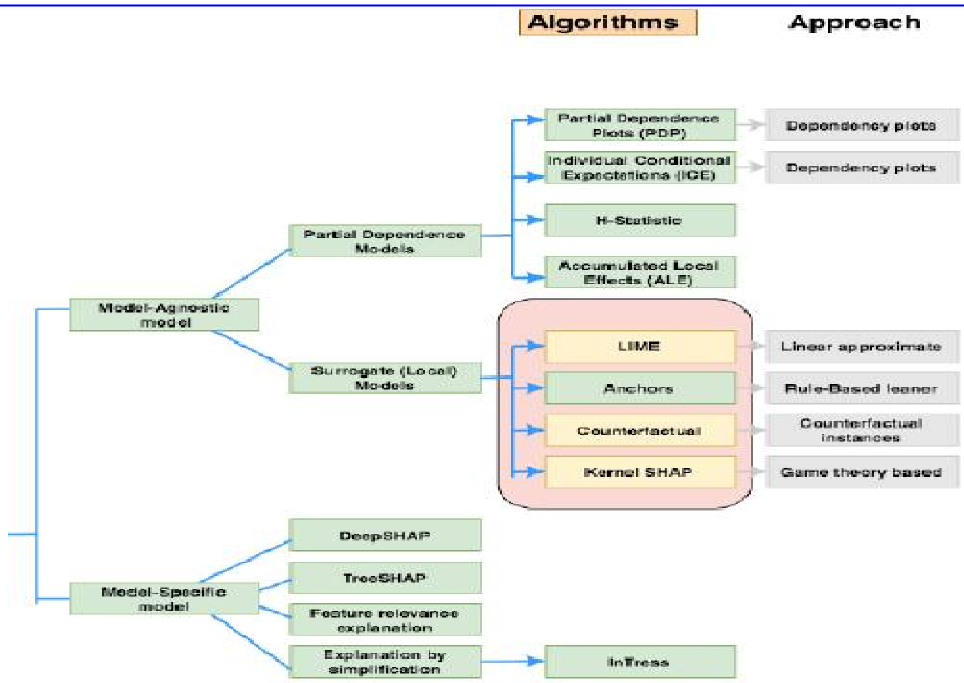
SWRL rules generated on the basis of 20 training samples using decision tree learning algorithm and special Protégé plugin



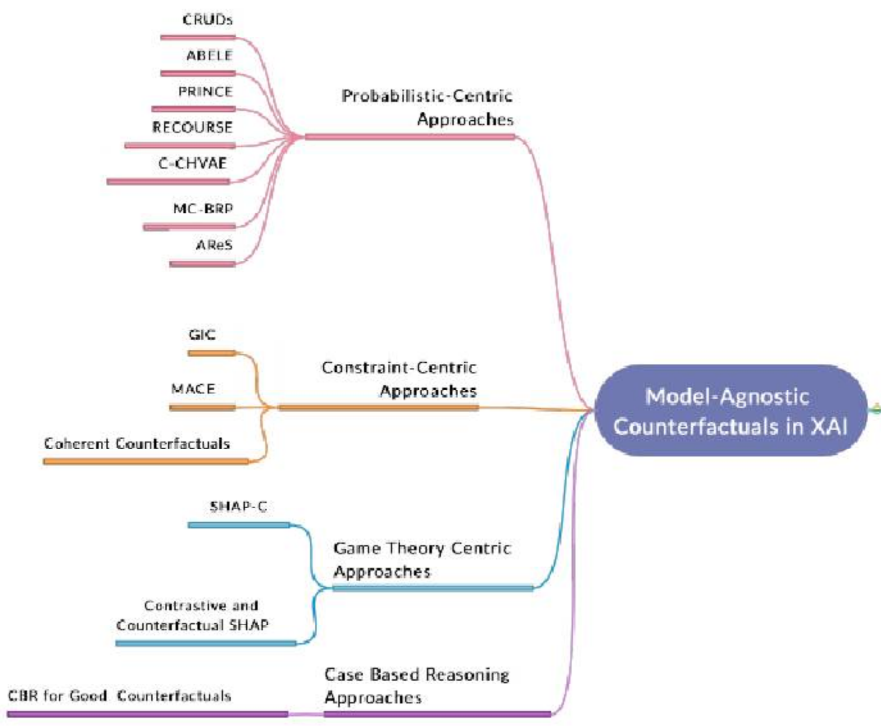
xAI.		2022-143

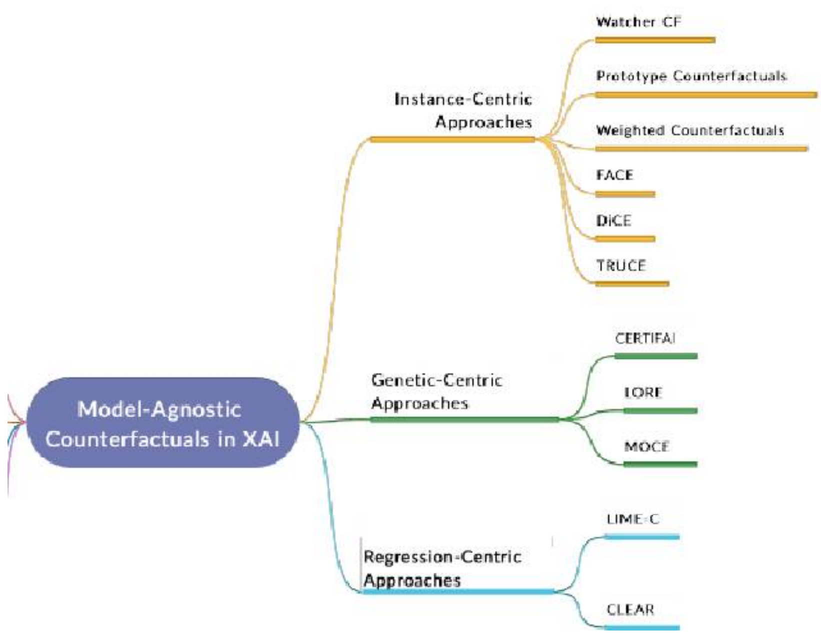
Taxonomy of explainable artificial intelligence


Taxonomy
Models
Explainable Models
Transparent Models
Opaque Models
Tree base model
K-Nearest Neighbour
Regression model
Rule-based models
Non-Gaussian Outcomes
Decision Tree
Bayesian network
Boosted Trees
Linear Regression
Logistic regression
Generalised Linear Models (GLM)
Generalised Additive Models (GAM)
Interaction Models
Random Forest
Support Vector Machines
Multi-layer Neural Network
Post-Hoc Explainability


Algorithms
Approach
Model-Agnostic model
Model-Specific model
Partial Dependence Models
Surrogate (Local) Models
Partial Dependence Plots (PDP)
Individual Conditional Expectations (ICE)
H-Statistic
Accumulated Local Effects (ALE)
Dependency plots
Dependency plots
LIME
Anchors
Counterfactual
Kernel SHAP
Linear approximate
Rule-Based leaner
Counterfactual instances
Game theory based
DeepSHAP
TreeSHAP
Feature relevance explanation
Explanation by simplification
inTrees

Proposed Taxonomy for model-agnostic counterfactual approaches for XAI

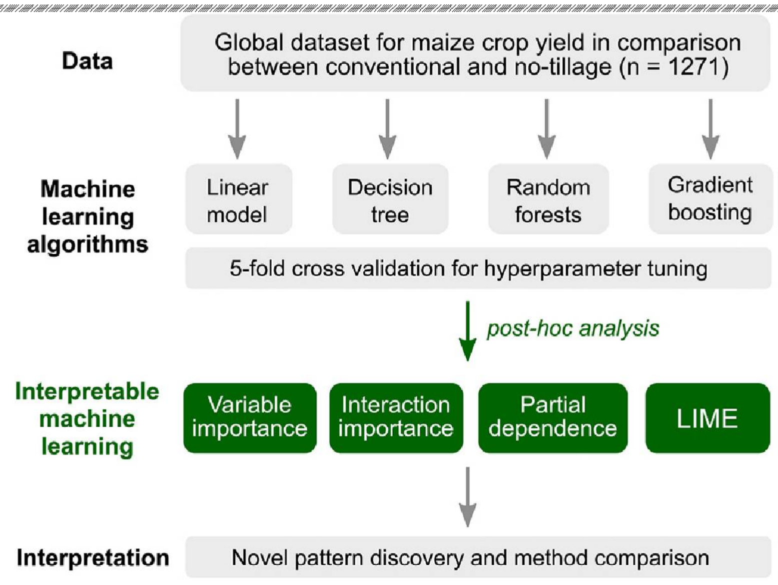




Classification of collected model-agnostic counterfactual algorithms for XAI based on different properties, theoreticalbackgrounds and applications

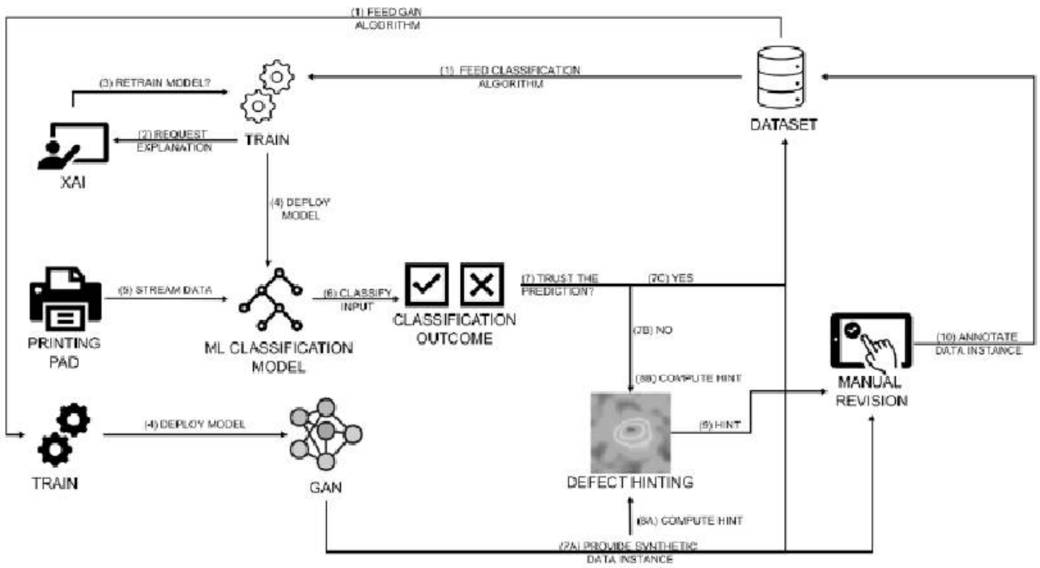
Theory / Approach	Algorithms	Ref.	Applications	Code?	Properties: Proximity	Plausibility	Sparsity	Diversity	Feasibility	Optimization	Causal?
Instance-Centric	WatcherCF	[55]	C [Tab / Img]	Yes [120] [Algo: CF]	✓ [L_1-norm]	✗	✓	✗	✗	Gradient Descent	✗
	Prototype Counterfactuals	[121]	C [Tab / Img]	Yes [120] [Algo: CFProto]	✓ [L_1/L_2-norm]	✓	✓ [kd-trees]	✓	✗	FISTA	✗
	FACE	[56]	C [Tab / Img]	Yes [122]	✗	✓	✗	✓	✓	ε-graphs	✗
	Weighted Counterfactual	[101]	C [Tab]	No	✓ [L_1-norm]	✗	✓	✗	✗	Gradient Descent	✗
	TRUCE	[102, 103, 104]	C [Tab / Txt / Img]	Yes [123]	✓ [L_1-norm]	✗	✓	✗	✗	Growing Spheres	✗
	DICE	[50]	C [Tab]	Yes [124]	✓ [L_1-norm]	✗	✓ [hinge loss]	✓	✓	Gradient Descent	✗
Probabilistic-Centric	CRUDS	[116]	C [Tab]	No	✓ [L_1-norm]	✓	✗ [Variation Autoencoders]	✓	✓	-	✗
	AReS [Global]	[125]	C/R [Tab/Txt]	No	✓ []	✗	✗ [Probabilistic]	✗	✗	Maximum a Posterior Estimate	✗
	PRINCE	[111]	C/R [Tab/Txt]	Yes [126]	✓	✗	✗ [Random Walk]	✗	✗	PageRank	✗
	C-CHVAE	[112, 113]	C [Tab]	Yes [127]	✓	✗	✓ [Variation Autoencoders]	✗	✗	Integer Programming Optimization	✗
	ABELE	[115]	C [Img]	Yes [128]	✓	✗	✓ [Variation Autoencoders]	✗	✗	-	✗
	RECOURSE	[117]	C [Tab]	Yes [129]	✓	✓	✓ [Variation Autoencoders]	✓	✓	Gradient Descent	✓
	MC-BRP	[114]	R [Tab]	Yes [130]	✓	✗	✓	✓	✗	Monte Carlo	✗
Constraint-Centric	GIC	[106]	C [Tab]	No	✓	✗	✓	✗	✗	Hill Climbing/ Genetic Algorithms	✗
	MACE	[93]	C [Tab]	Yes [131]	✓ [$L_0/L_1/L_\infty$-norm]	✓	✓ [constraint satisfaction]	✓	✓	SMT	✗
	Coherent Counterfactuals	[98]	C / R [Tab / Txt]	Yes [132]	✓ [L_1-norm]	✓	✓ [mixed polytopes]	✓	✓	Gurobi Optimization	✗
Genetic-Centric	MOCE	[92]	C [Tab]	Yes [133]	✓ [L_1-norm]	✗	✓ [min feature changes]	✗	✗	NSGA-II	✗
	CERTIFAI	[107]	C [Tab / Img]	Yes [134]	✓ [L_1-norm / SSIM]	✗	✗	✓ [mutations]	✗	Fitness	✗
	LORE	[108]	C [Tab]	Yes [135]	✓ [L_2-norm / Match]	✗	✗	✓ [mutations]	✗	Decision Tree Model	✗
Regression-Centric	LIME-C	[136, 109]	C / R [Tab / Txt / Img]	Yes [137]	✗	✗	✗	✗	✗	Additive Feature Attribution	✗
	SED-C	[95]	C [Txt]	Yes [138]	✗ [cosine similarity]	✗	✗	✗	✗	-	✗
	CLEAR	[108]	C [Tab]	Yes [139]	✓ [L_1-norm]	✗	✓ [min feature changes]	✗	✗	Regression	✗
Game Theory Centric	SHAP-C	[136, 109]	C / R [Tab / Txt / Img]	Yes [140]	✗	✗	✗	✗	✗	Shapley Values	✗
	SHAP-CC	[118]	C / R [Tab]	No	✗	✗	✗	✗	✗	Shapley Values	✗
Case Based Reasoning	CBR for Good Counterfactuals	[96]	C [Tab / Txt]	No	✓ [L_1-norm]	✓	✓ [counterfactual potential]	✓	✓	Nearest Unlikely Neighbour	✗

Post-hoc analysis



xAI.		2022-166

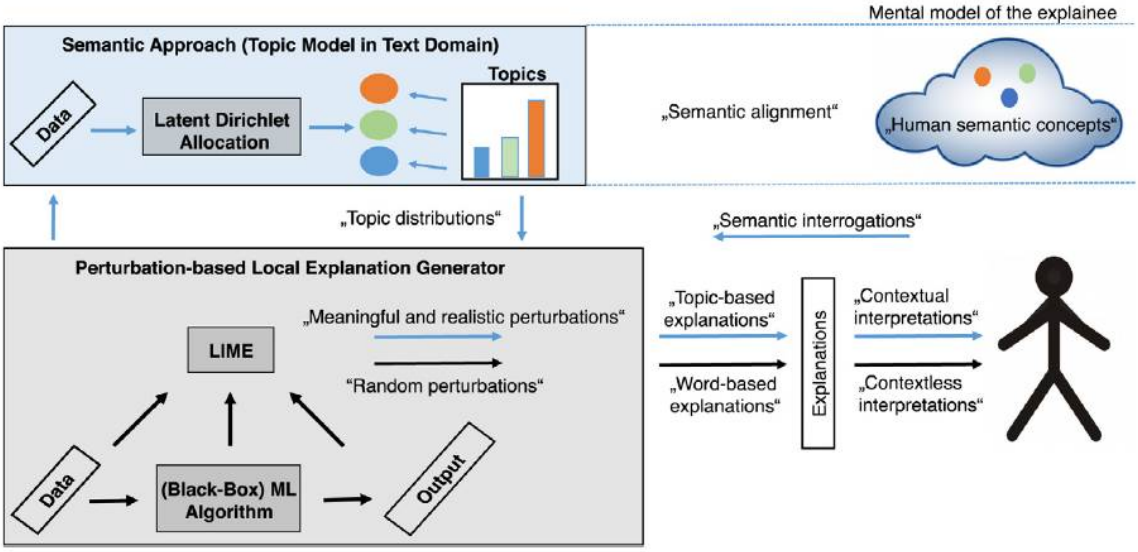
Visual quality inspection pipeline



Non-defective (good)
Defective(double and interrupted) prints

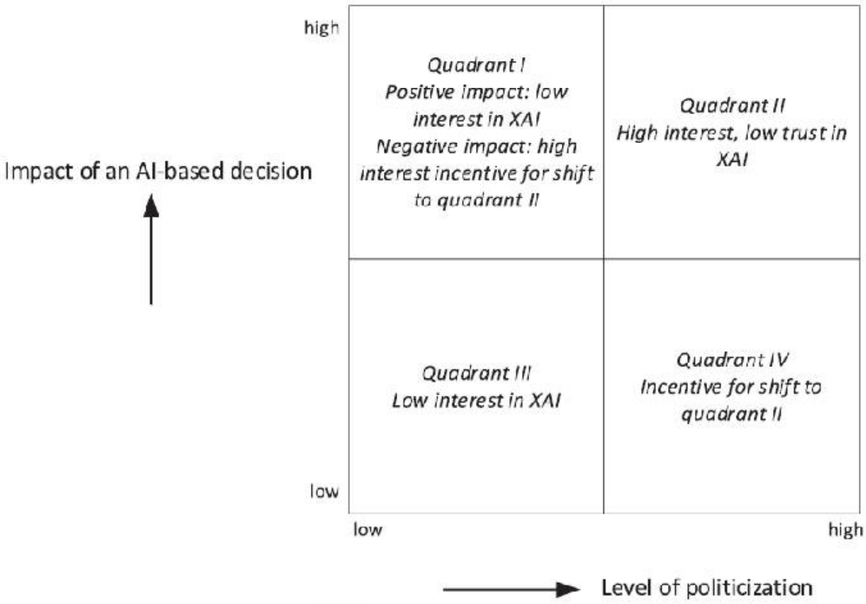
xAI.		2022-167

Integration of an ML classification algorithm
with LIME and semantic approach



- ✓ Black arrows:Classical way of generating and communicating explanations in a model-agnostic and perturbation-based way
- ✓ Blue arrows: Explanation process of CaSE integrating a semantic approach

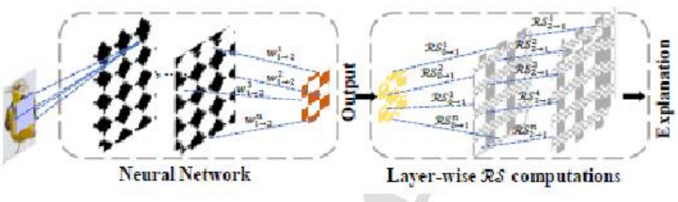
xAI.	Typical challenges of XAI.	2022-173



Challenge	Explanation
1. Lack of expertise	Most persons will lack the expertise to understand the explanation and assess the fairness of the decision.
2. Contested explanations	Experts explaining algorithms also make biased and inherently disputable choices.
3. Dynamics of data and decisions	Data and decisions change over time, and therefore explanations change.
4. Interference of algorithms	Often there is a whole chain of activities to collect and process data from various types of sources, and many, often different kinds of algorithms are used.
5. Context-dependency	Algorithms cannot be explained at a general level, as outcomes might be different per individual.
6. Wicked nature of the problems addressed	Wicked problems are ill-structured, are ambiguous by nature and can be solved in different ways. Algorithms provide one answer that is contestable and changes over time.
7. Causality is not used for making decisions	If the causality is explained between inputs and outputs, this does not mean that the algorithm uses that causality to arrive at a decision. Furthermore, the explanation of causality might change over time.

xAI. 2023-030

LRP

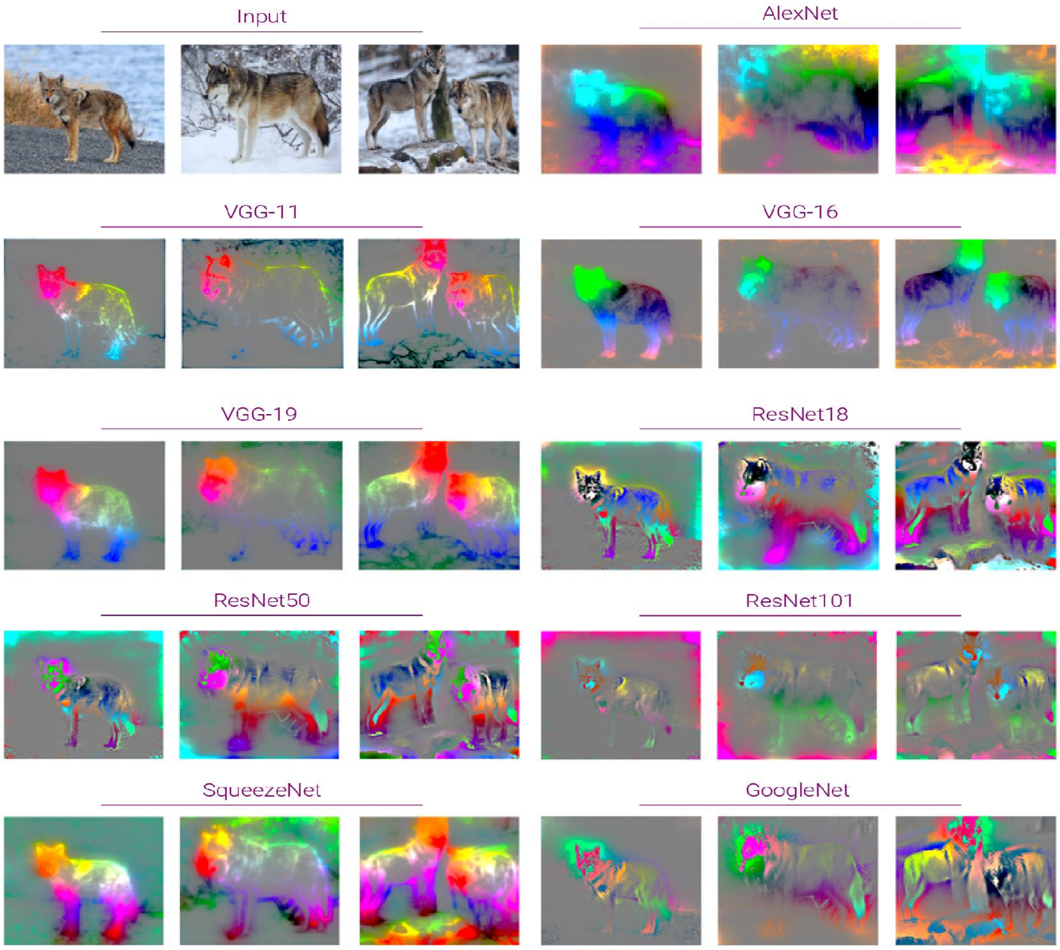


- ✓ LRP: Decomposes a model's prediction function into a sum of layer-by-layer relevance values.
- ✓ In other words, it is a Deep Talyor Decomposition of a prediction when used with ReLU networks

PRincipal Image Sections Mapping PRISM

Feature indication potential of PRISM

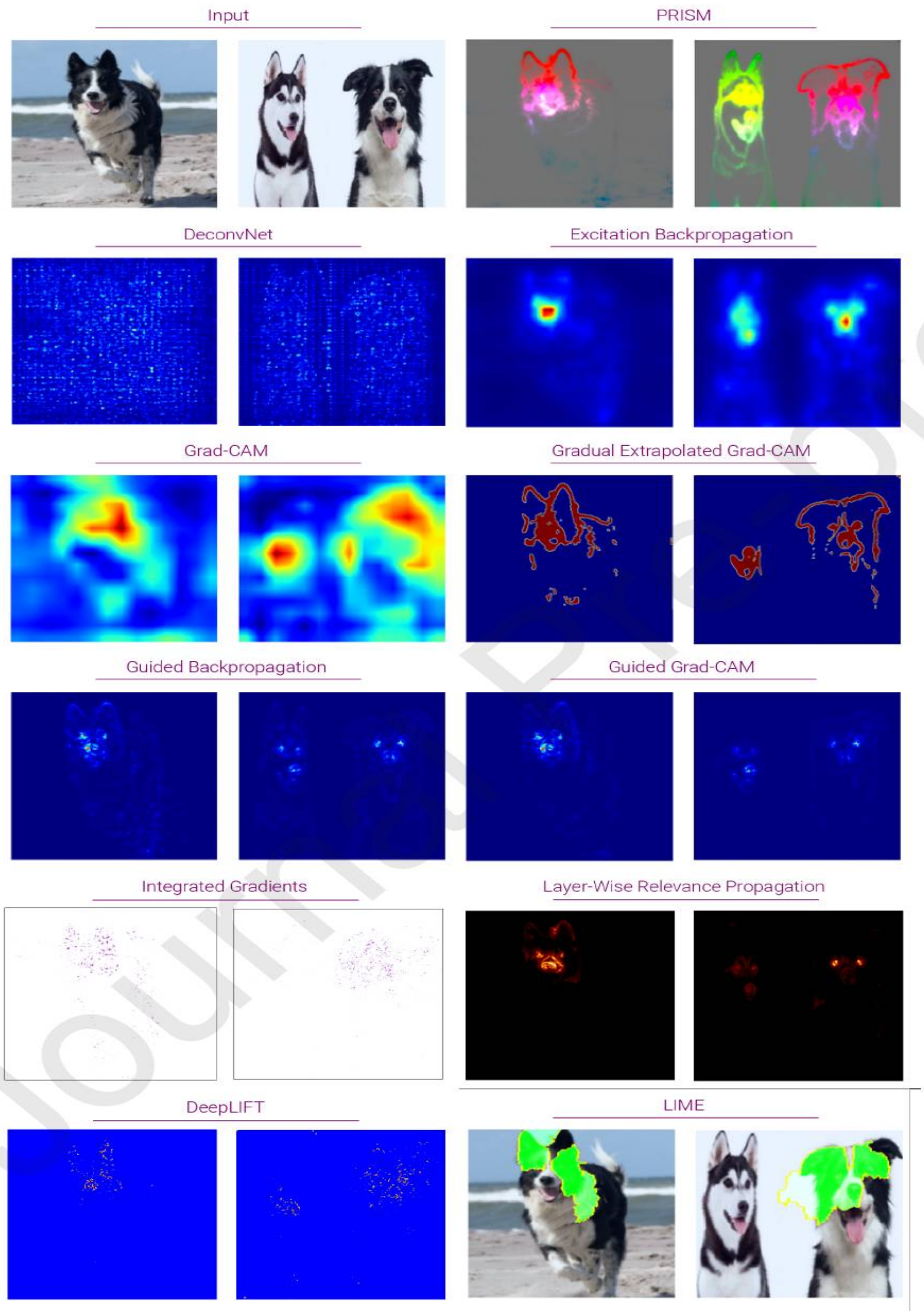
Images of two wolves and one coyote



✓ Two images of wolves and one of a coyote were used to depict the feature indication potential of PRISM

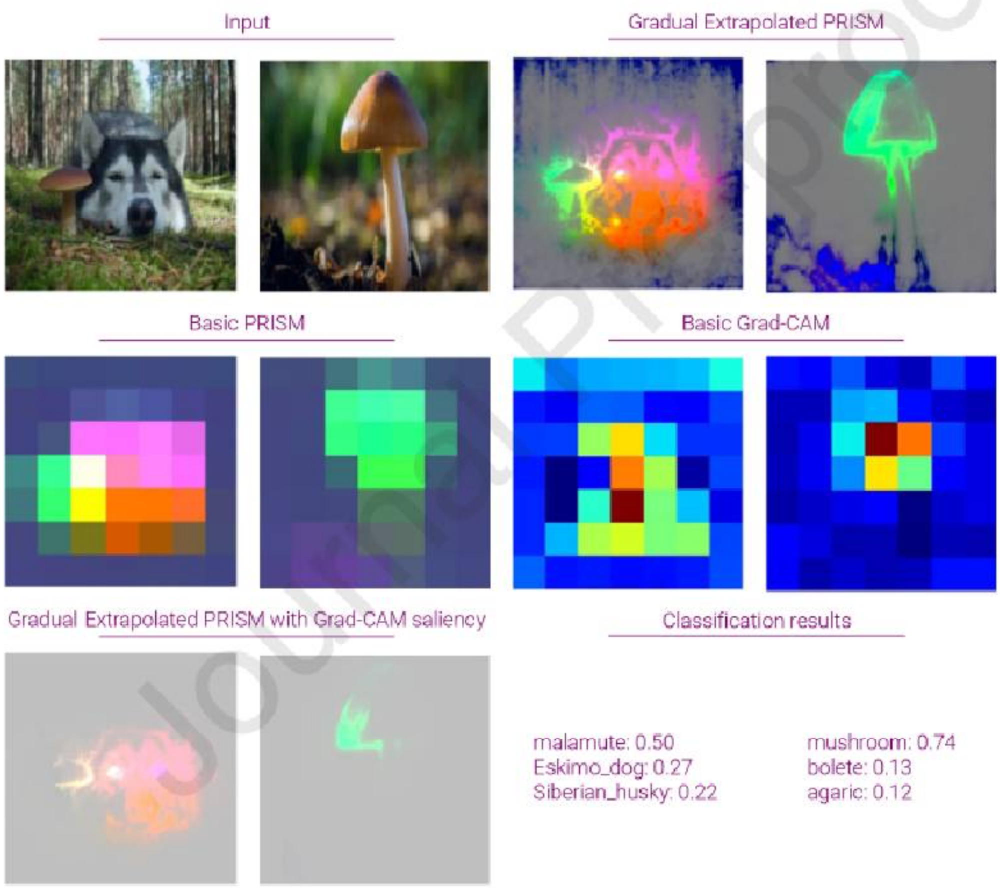
+ PRISM outputs for multiple state-of-the-art models

Output comparison of PRISM



xAI.		2023-061

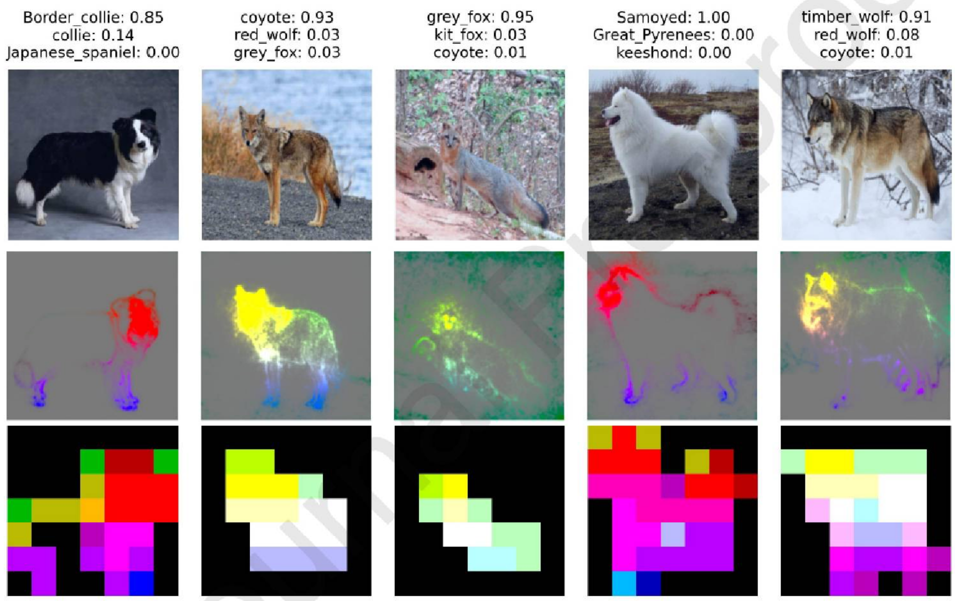
PRISM detects features (mushroom) ignored by final CNN's verdict



✓ Addition of salience provides a proper result

xAI. | 2023-061

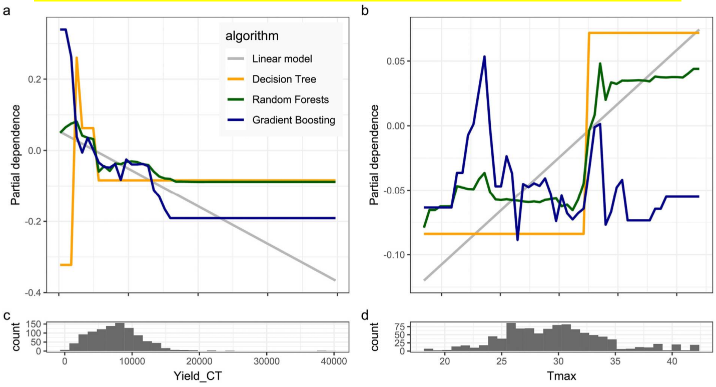
PRISM for five canine classes



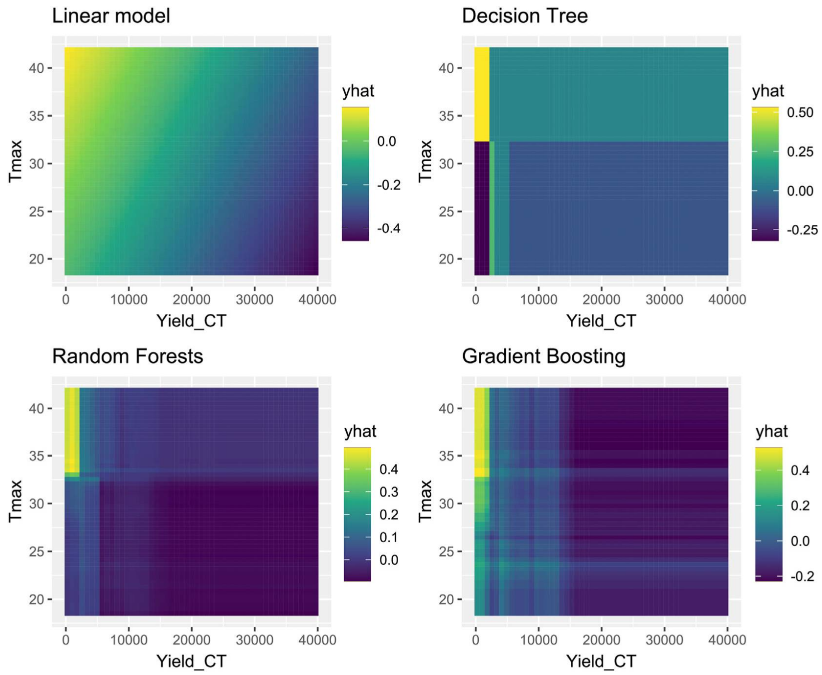
First row: top three confidence scores for each image
Second row: input images
Third row: GE PRISM
Last one: PRISM with only exclusive colours

xAI.	Agricultural science	2022-153

Partial dependence plots for
Yield_CT (a); Tmax (b) with the data distributions (c, d)



Partial dependence plot (2D)

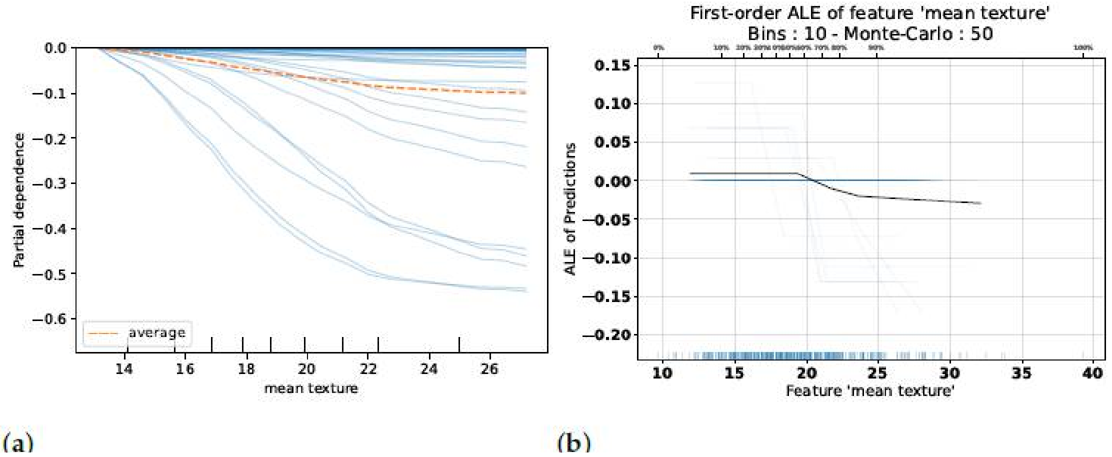


A brighter yellowregion (top-left) indicates that crop yield in no-tillage is higher than conventional tillage,while a darker blue region (bottom-right)indicates the opposite

Variable importance plots

xAI.		2023-

(a) PDP & ICE curves(b) ALE curves

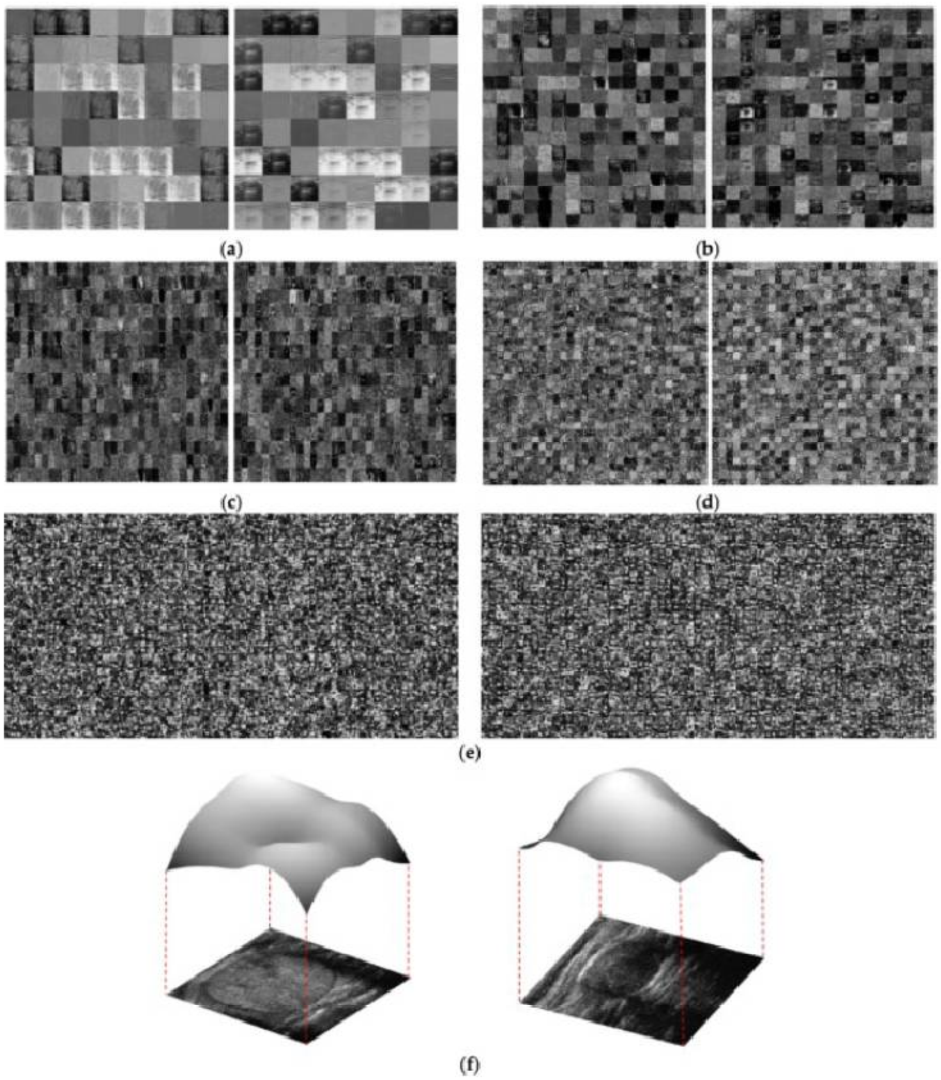


Feature activation maps

xAI.		2022-066

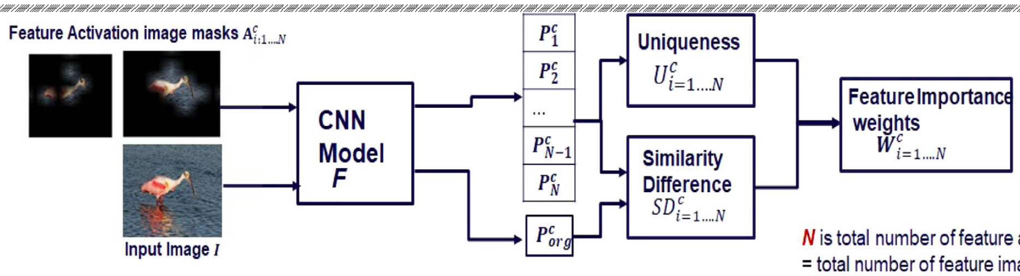
Feature activation maps

Thyriod Nodule classification – ultrasound Data



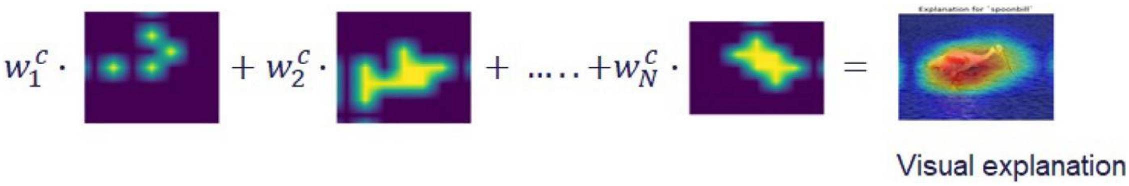
✓ Feature activation maps for two input images from the 1st (a), 10th (b), 22nd (c), 40th (d) and 49th (e) convolutional layer of ResNet

xAI.		2022-124
	Feature activation image mask A	



N is total number of feature activations = total number of feature image masks
P_i is the predictions of feature image masks
P_{org} is the predictions of original input image masks

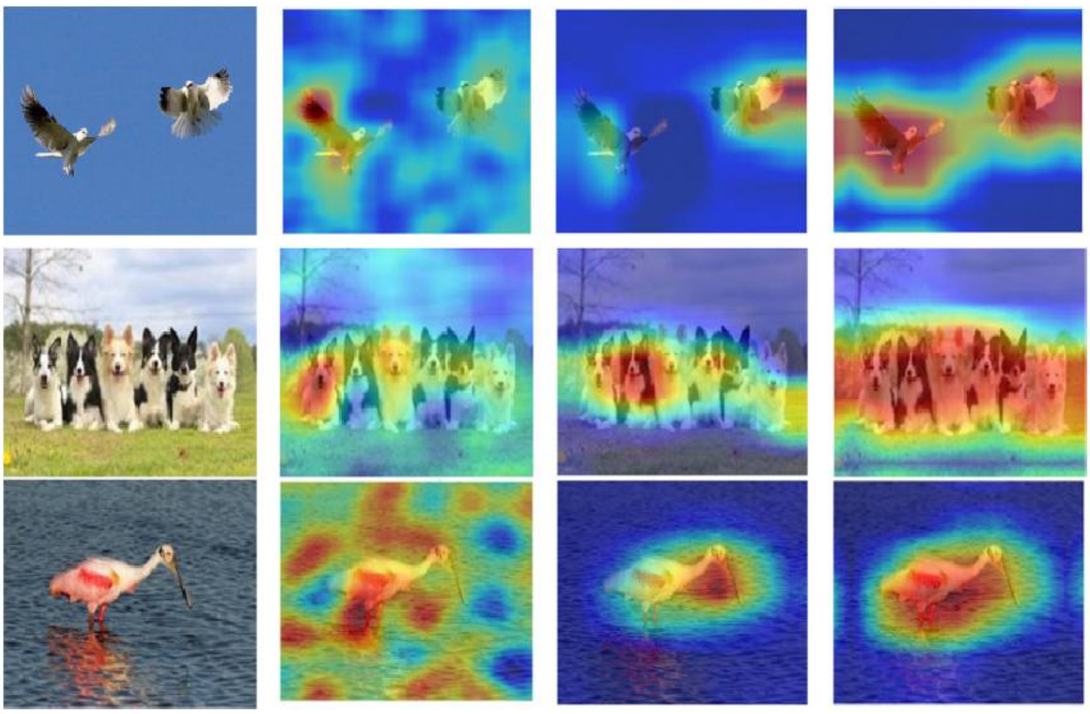
Visual explanation for prediction



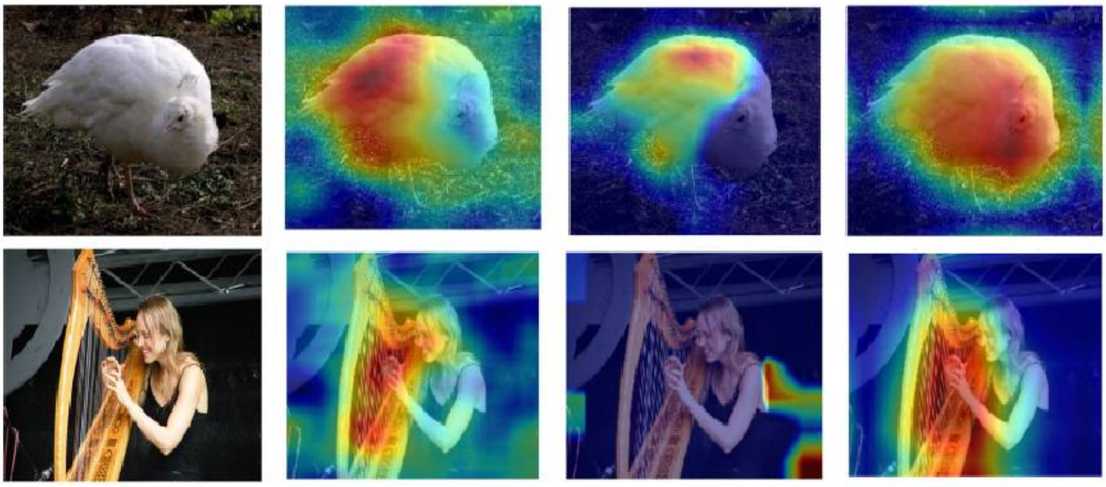
✓ Visual explanation is a weighted linear combinations of feature activation masks for prediction of the class

xAI.	2022-124

Visual comparison of explanation maps generated for natural images classes predicted by CNN model 'Bird','Borzoi dog', 'Spoonbill', 'Goose', and 'Harp'

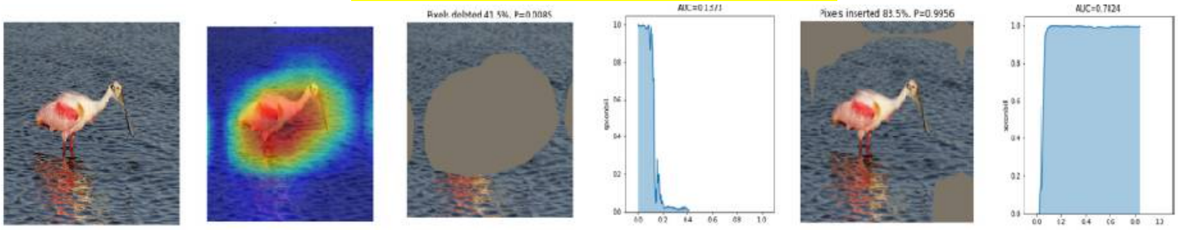


(a) Original Image (b) RISE (c) GRAD-CAM (d) SIDU



(a) Original Image (b) RISE (c) GRAD-CAM (d) SIDU

Insertion and deletion casual metrics AUC

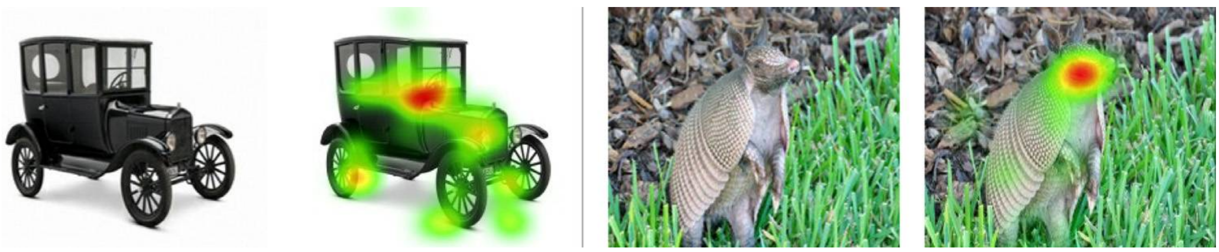


(a) (b) (c) (d) (e) (f)

- ✓ (a) original image (b) SIDU explanation map
- ✓ (c) deletion metric (d). (e) insertion metric (f) AUC

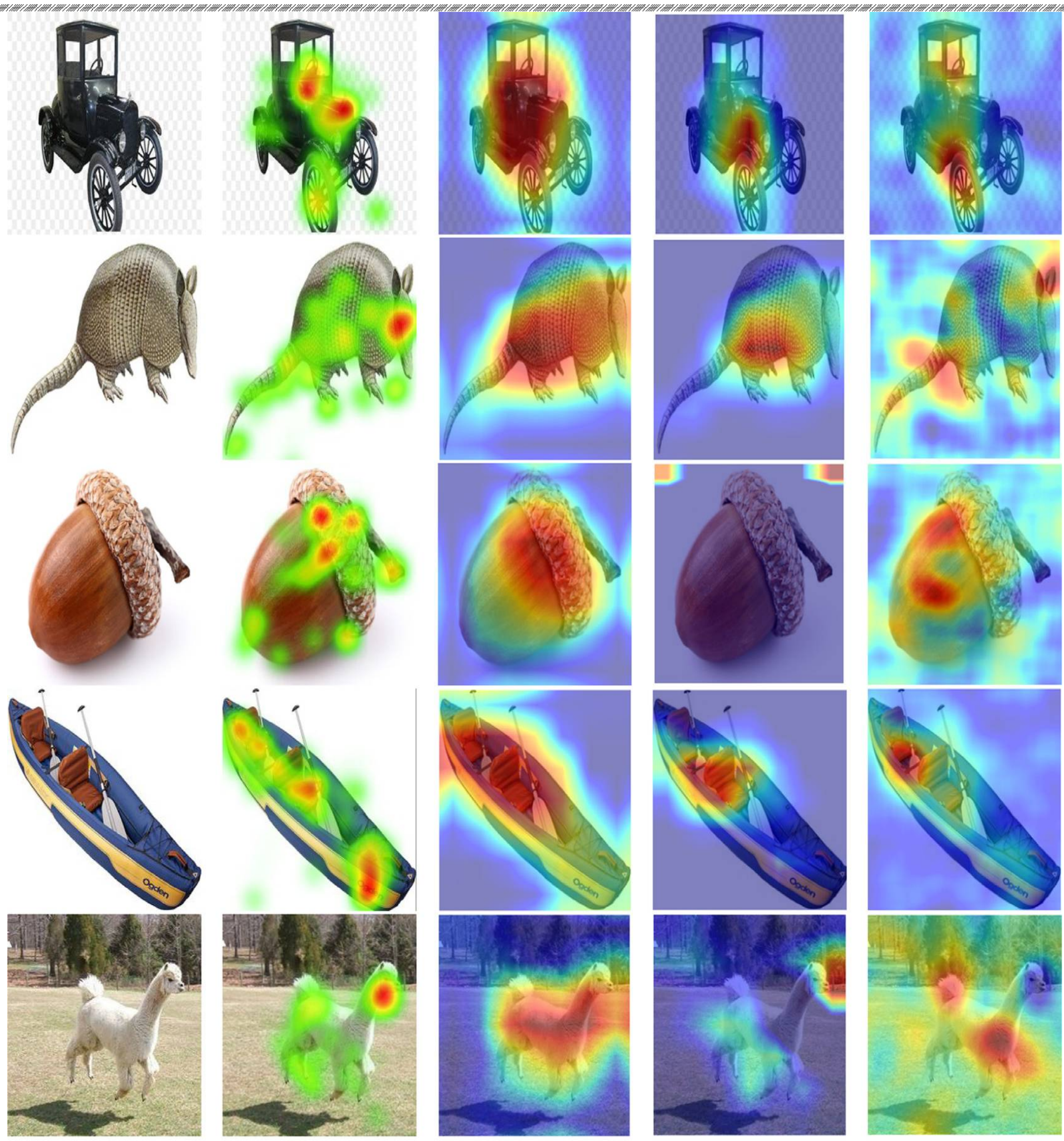
xAI.		2022-124

Examples of Eye-tracking data collection from humans
to recognize given object classes ‘Model T and ‘Armadillo

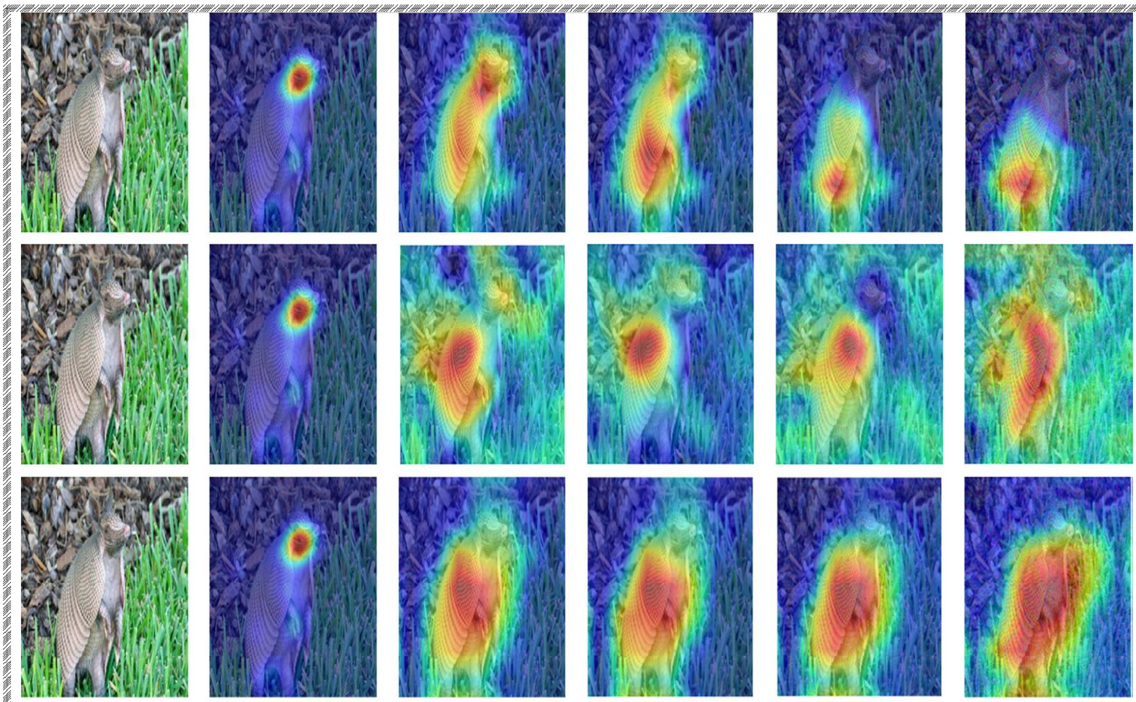


xAI.		2022-124

Comparison of XAI methods
Visual explanation of object classes



Comparison of XAI visual explanation with different levels of FGSM noise
with human visual explanation (heatmaps)

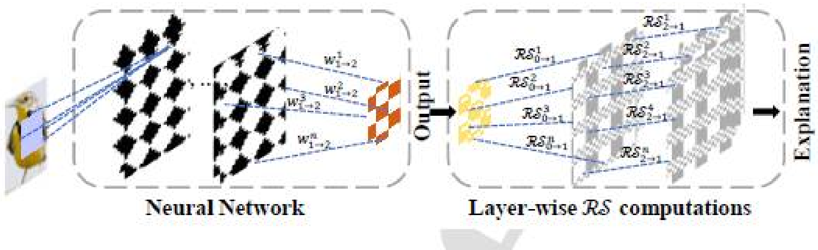


(a) (b) (c) (d) (e) (f)

LRP

xAI.		2023-

LRP



- LRP: Decomposes a model's prediction function into a sum of layer-by-layer relevance values.
- LRP can be thought of as the Deep Talyor Decomposition of a prediction when used with ReLU networks.

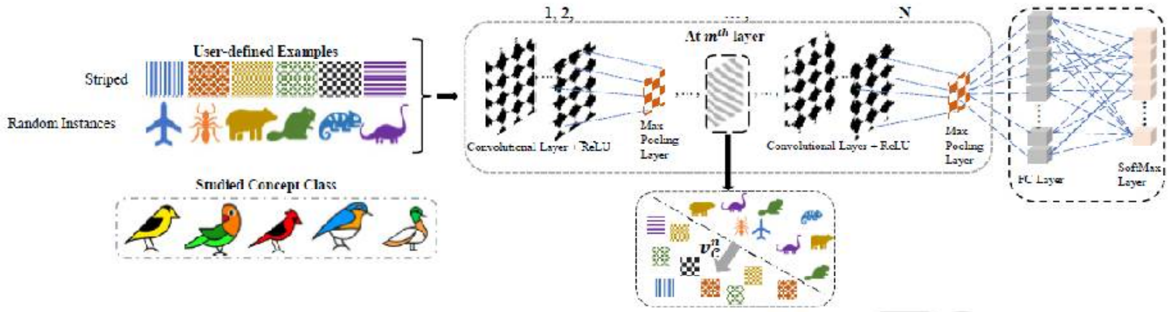
xAI.		2023-150

List of XAI studies that used LRP to explain their model prediction results

Author	Objective	Subject	Application	Data type	Data	ML/DL	Classifier	Results
Binder et al. [52]	Morphological and molecular breast cancer profiling	565 BC	CDS	Image	Histological image	ML	SVM	ACC: 98.00%
Chereda et al. [54]	patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer	393 metastasis 576 no metastasis	Precision medicine	Genomic	gene expression data	DL	Graph CNN	ACC: 76.00% AUROC: 0.820
Böhle et	Alzheimer's Disease	193 AD	CDS	Image	MRI	DL	CNN	ACC: 87.96%

xAI.		2023-

Concept Activation Vectors (CAVs)



SHAP

xAI.		2023-142

Libraries for Shap and Lime

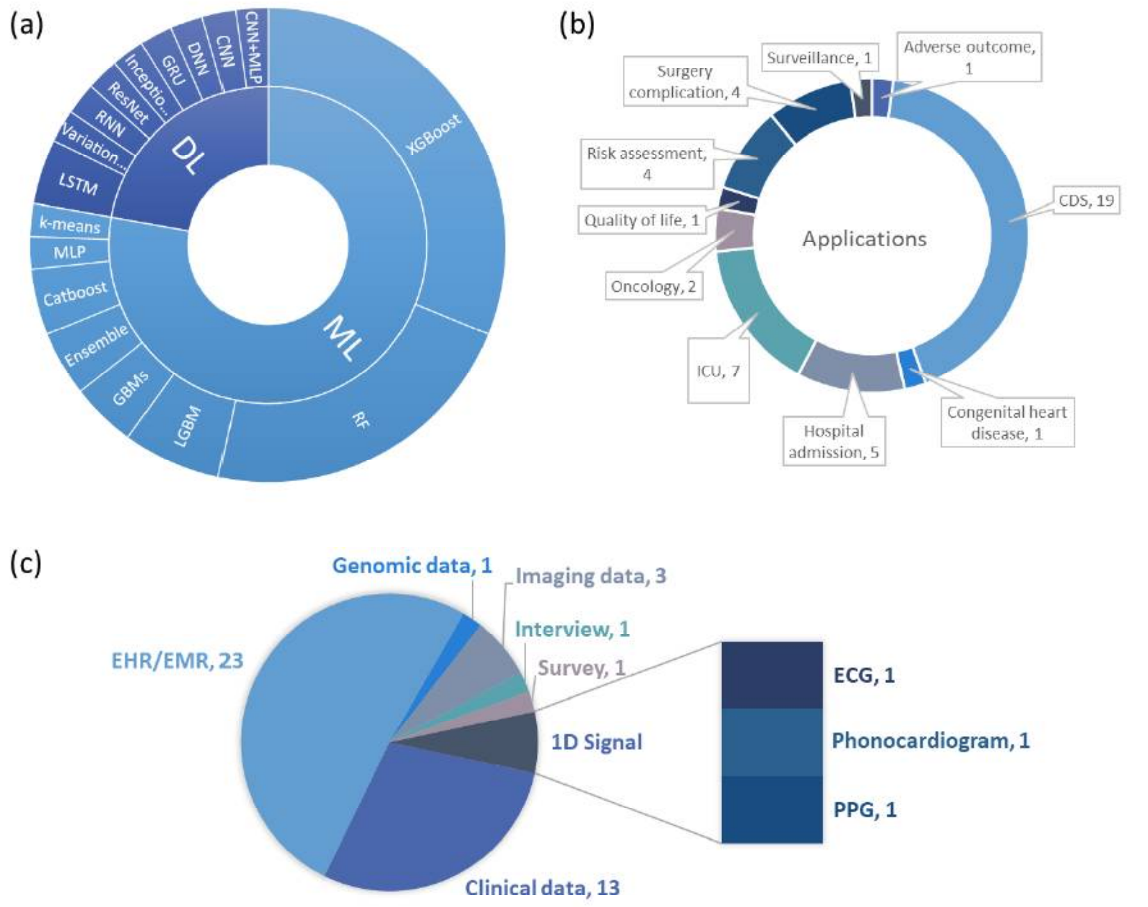
Library Name	Description	Library Version
NumPy	A library that implements linear algebra operations, mathematical functions, elements of statistical analysis	1.21.0
Matplotlib	Library for plotting various types of graphs	3.5.1
Scipy	Library designed to perform scientific and engineering calculations	1.8.0
Pandas	Library for working with tabular data structures	1.4.1
Shap	Library with implementation of the XAI SHAP method	0.40.0
Lime	Library with the implementation of the XAI LIME method	0.2.0.1
Scikit-learn	Library with tools for designing and training models	1.0.2

SHAP. MedData (MD) →xAIM

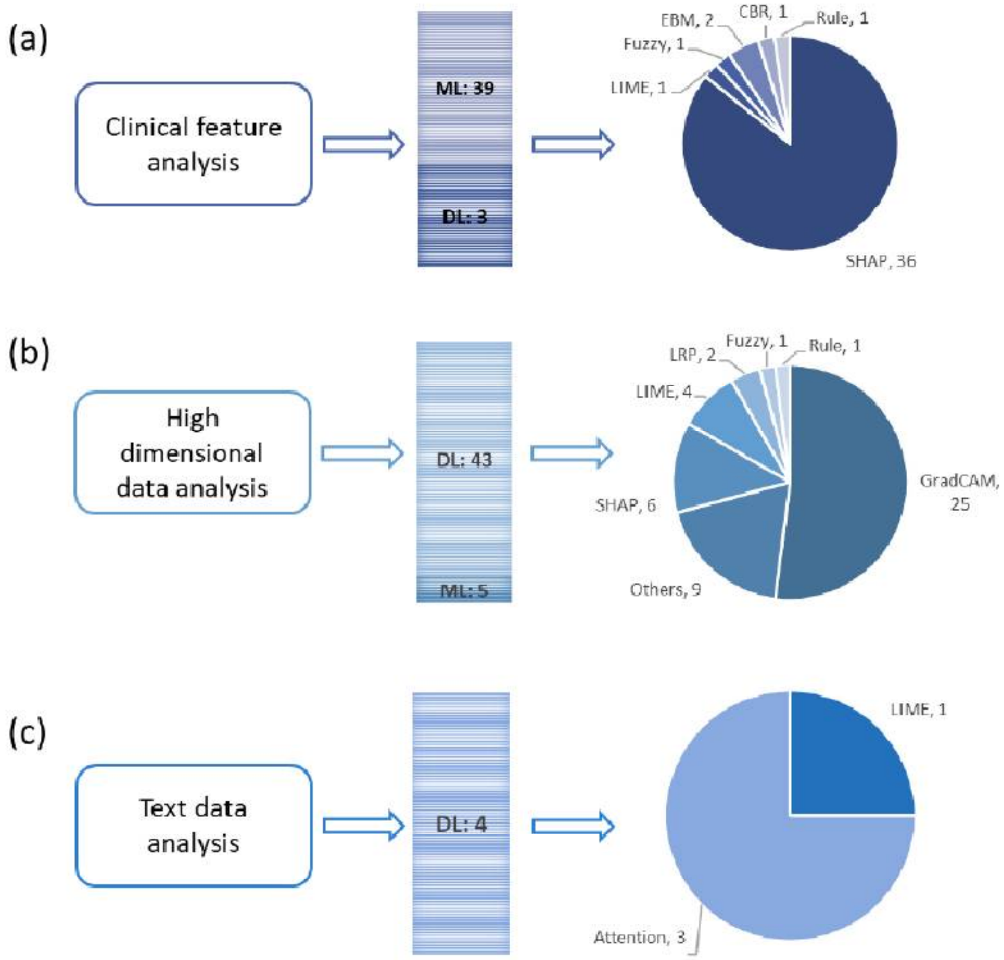
xAI.	Health care	2023-150

SHAPdiagram of AI models

Sunburst, Doughnut, Pie



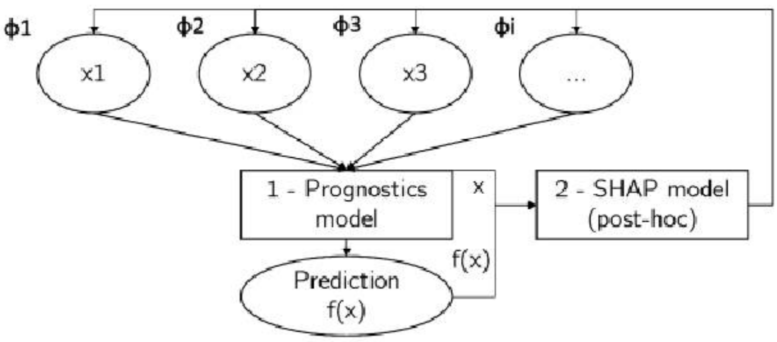
xAI.	Health care	2023-
AI models employed and the respective XAI technique		



✓ (a) clinical feature analysis, (b) high dimensional data analysis, (c) text data analysiss

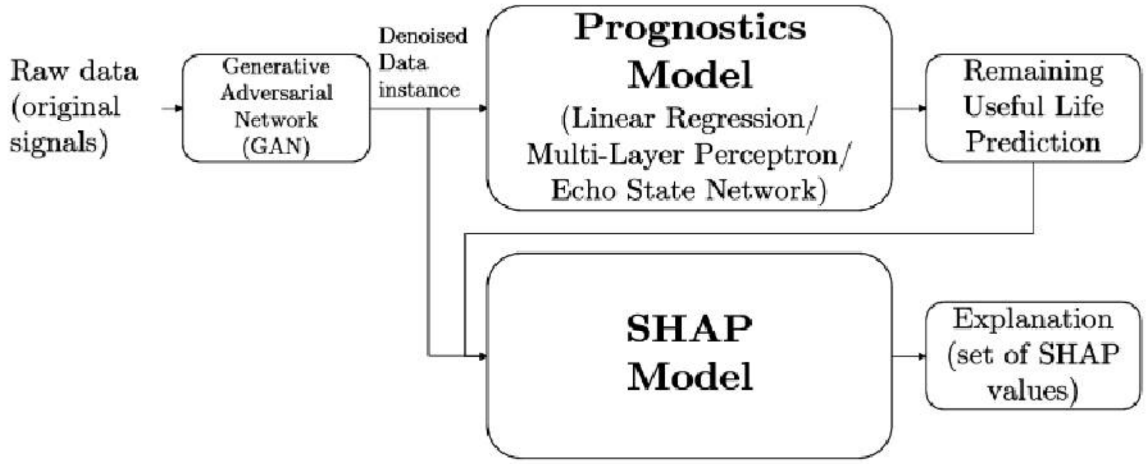
xAI.		2022-127

Input vector → prediction
Architecture – xAI -- model



- Black box model: inputs lead to the prediction without explicit causal relationships
+ SHAP model: Able to output decomposition factors (SHAP values) facilitating understanding of

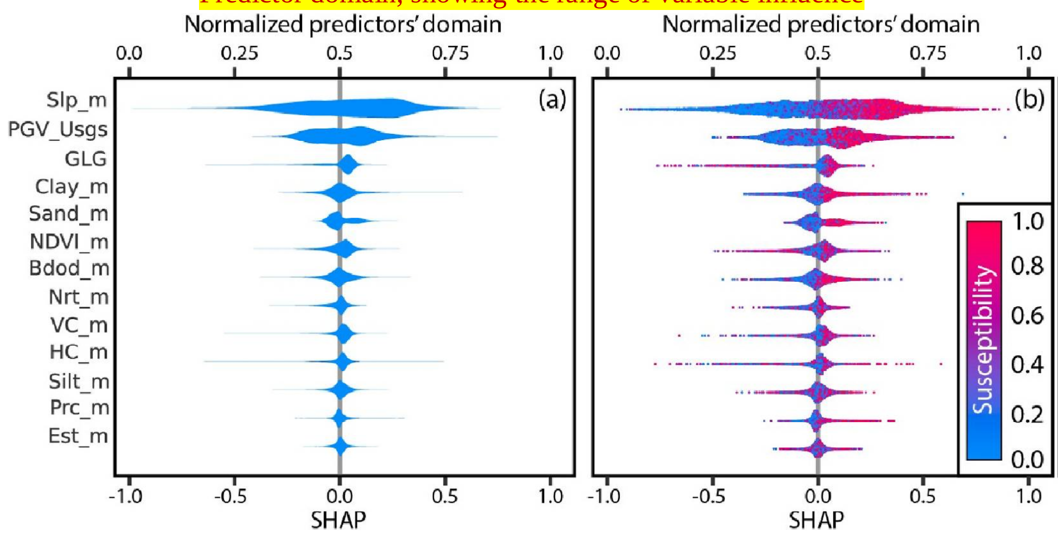
importance of each feature value to the prediction



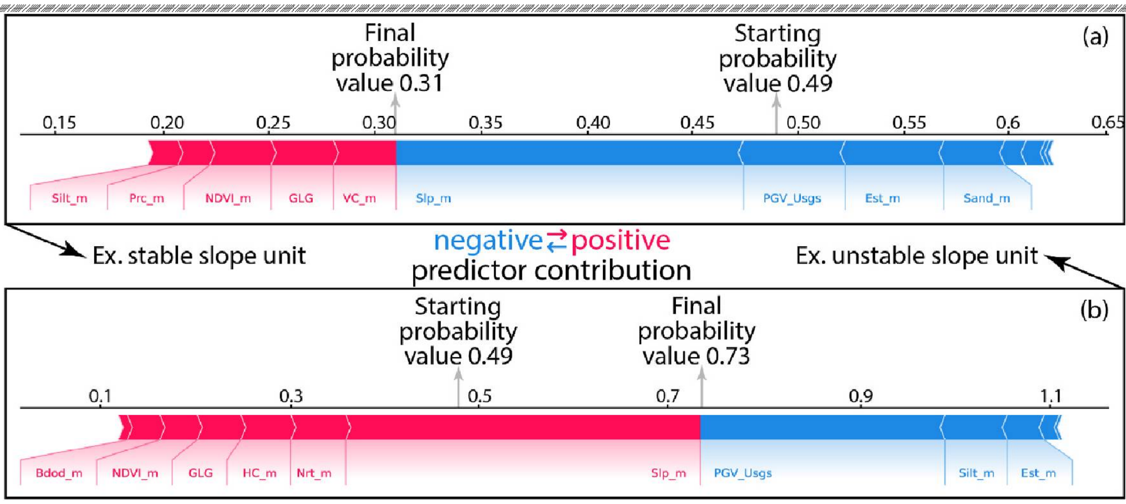
SHAP. GeoScience→xAI.GeoSci

xAI.	Geoscience, landslide susceptibility	2023-002

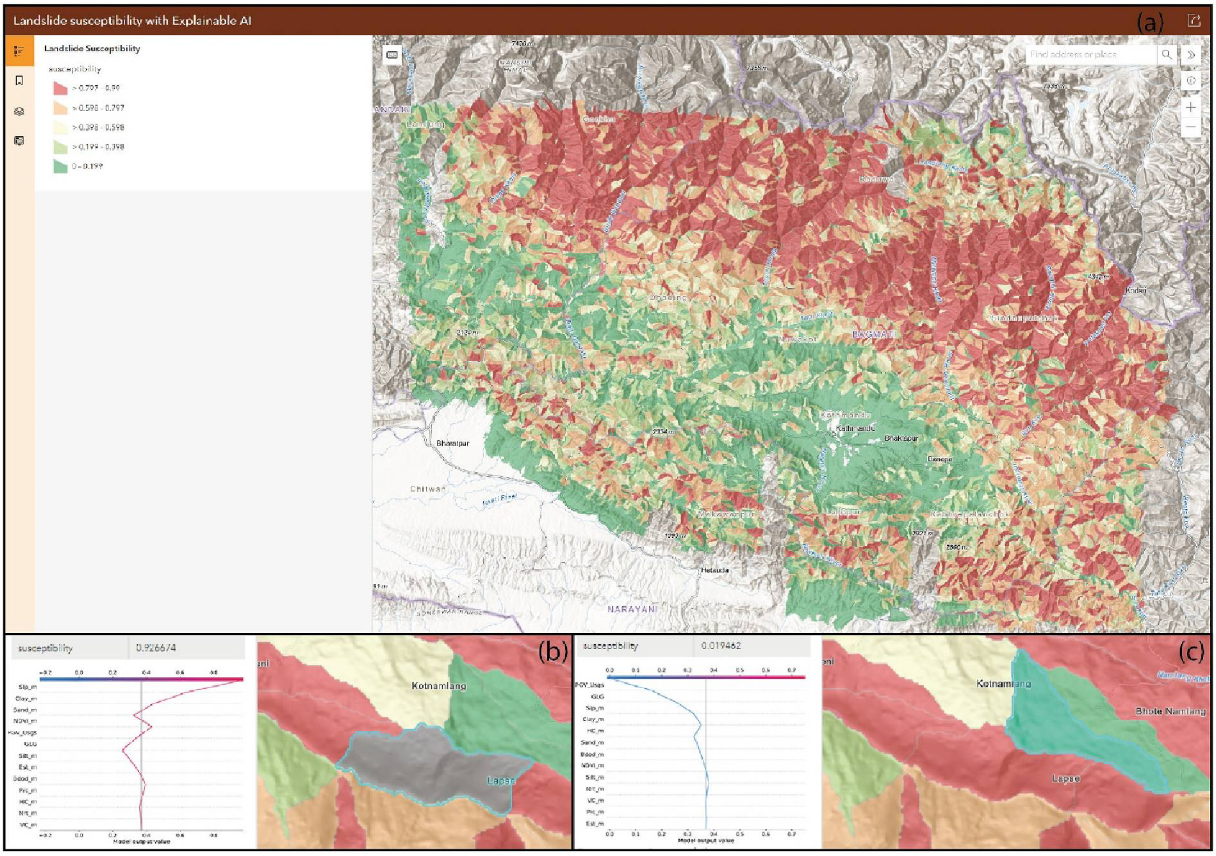
Predictor domain, showing the range of variable influence



Model's decision process to generate a given susceptibility value



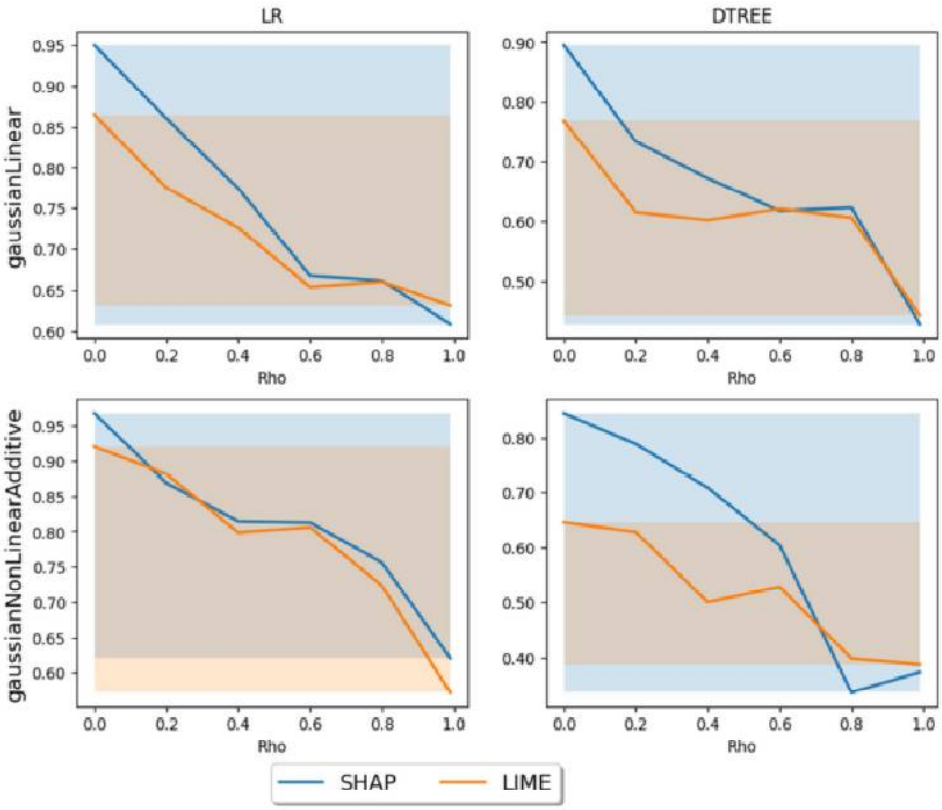
General overview of the web application



- ✓ Panel (a) : susceptibility map obtained by using our XAI
  - o Depicted here into five equal spaced classes
- ✓ Panel (b) : example of an XAI query for an unstable SU.
- ✓ Panel (c) : example of an XAI query for an stable SU.

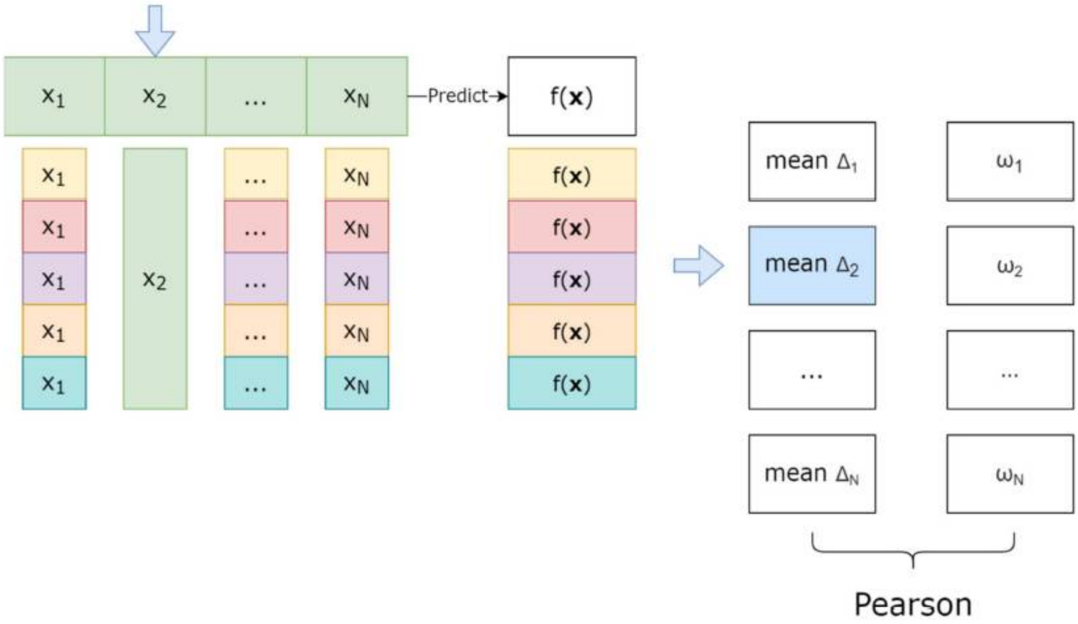
xAI.		2023-142

Dependence of the reliability metric on correlation coefficient



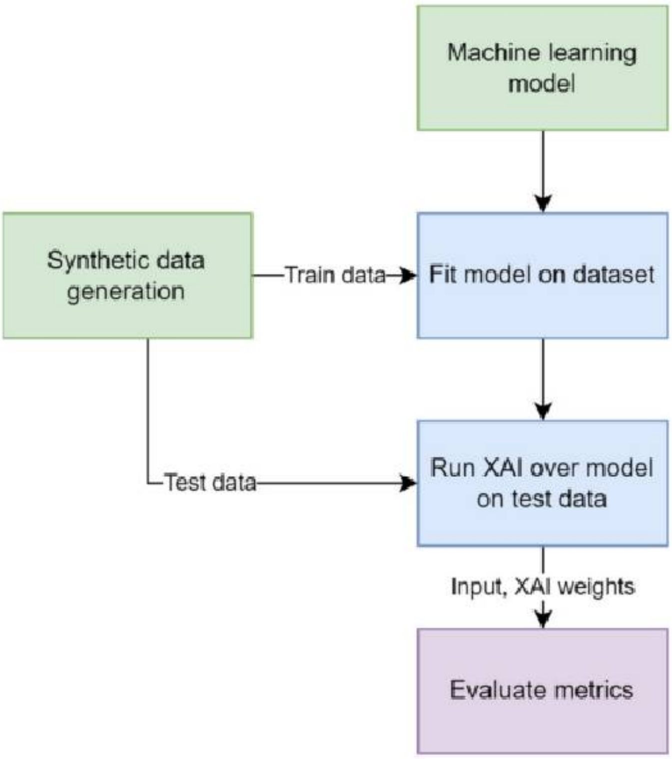
xAI.		2023-142

Algorithm for calculating faithfulness metric

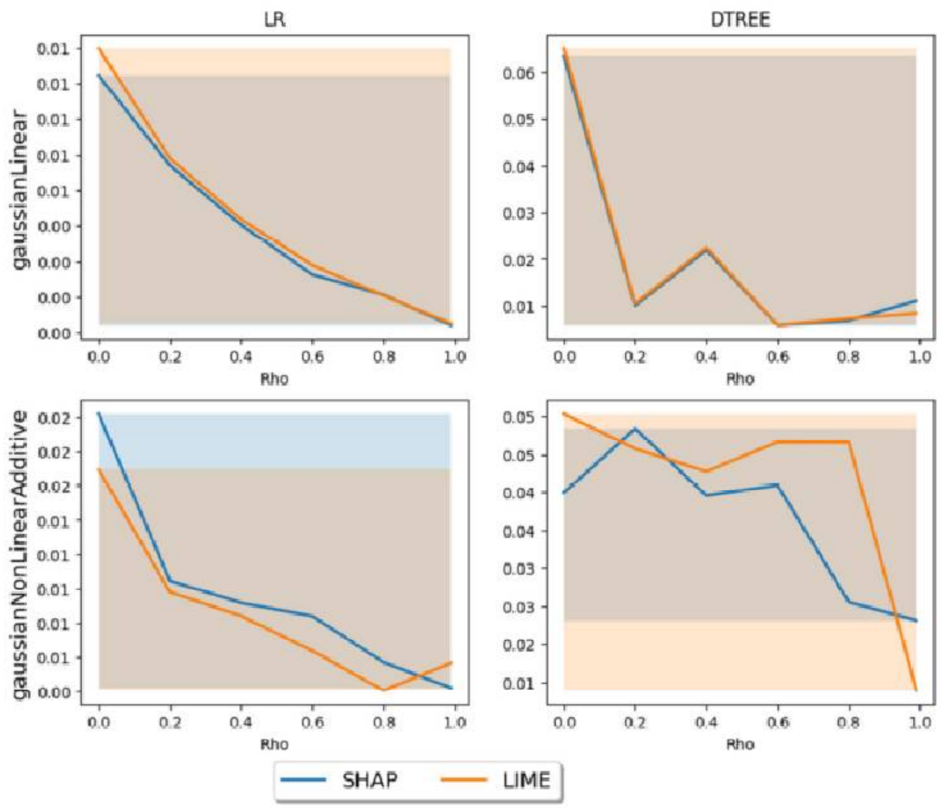


xAI.		2023-142

Methods flow

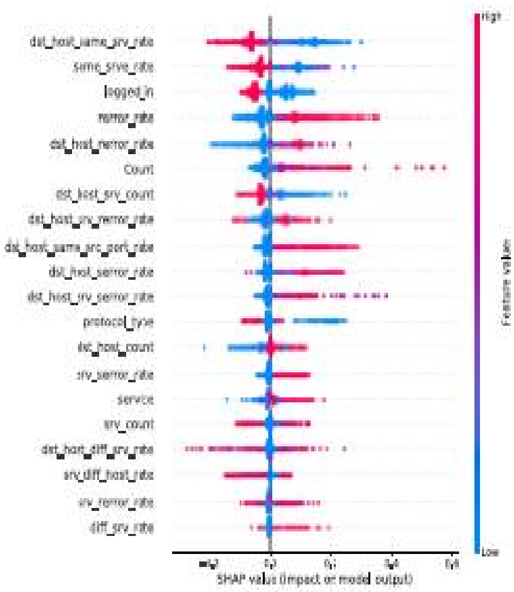


Incompleteness metric



SHAP; IoT

xAI.	Intrusion detectionframework in IoT networks	2023-057
Top 20 relevant features of attacks that binary classifiers learned SHAP		

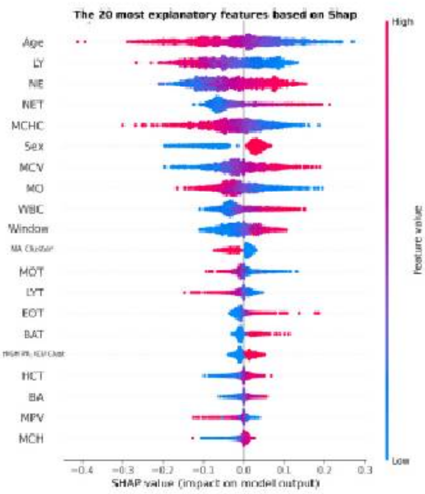


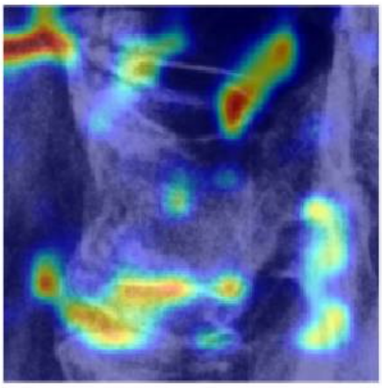
Weight	Feature
0.0587 ± 0.0023	dst_host_srv_count
0.0420 ± 0.0013	protocol_type
0.0217 ± 0.0011	logged_in
0.0203 ± 0.0007	dst_host_serror_rate
0.0140 ± 0.0007	hot
0.0101 ± 0.0012	service
0.0097 ± 0.0010	srv_serror_rate
0.0090 ± 0.0010	dst_host_same_src_port_rate
0.0084 ± 0.0014	dst_host_srv_rerror_rate
0.0078 ± 0.0004	dst_host_srv_serror_rate
0.0067 ± 0.0014	dst_host_rerror_rate
0.0064 ± 0.0009	dst_host_count
0.0052 ± 0.0025	dst_host_same_srv_rate
0.0049 ± 0.0008	same_srve_rate
0.0041 ± 0.0019	Count
0.0026 ± 0.0009	serror_rate
0.0008 ± 0.0002	root_shell
0.0007 ± 0.0004	is_guest_login
0.0007 ± 0.0017	rerror_rate
0.0002 ± 0.0001	num_shells
… 21 more …	

(a) NSL-KDD features extracted by SHAP (b) NSL-KDD features extracted by PFI

xAI.		2023-116

Examples of explanations commonly used in XAI

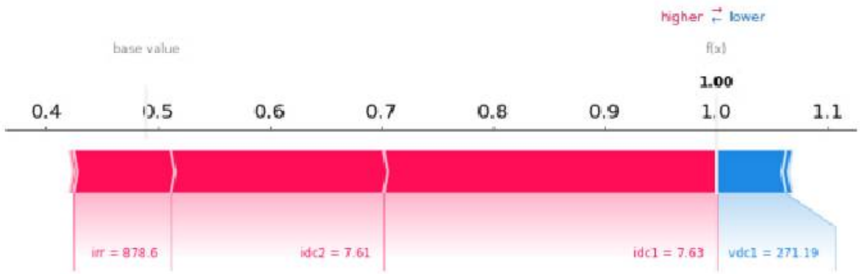




SHAP instance based	SHAP: a feature based activation map

xAI.	Fault detection	2023-136

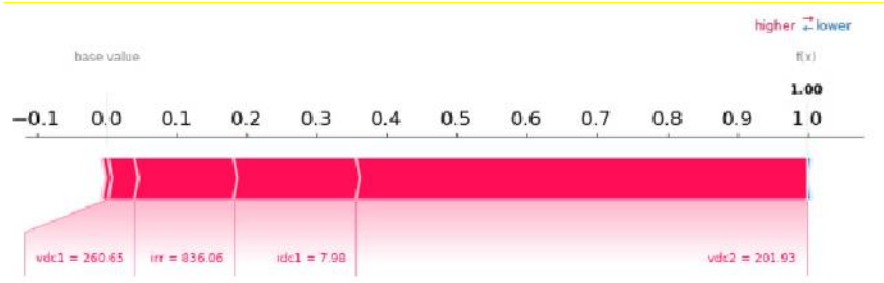
Healthy observation



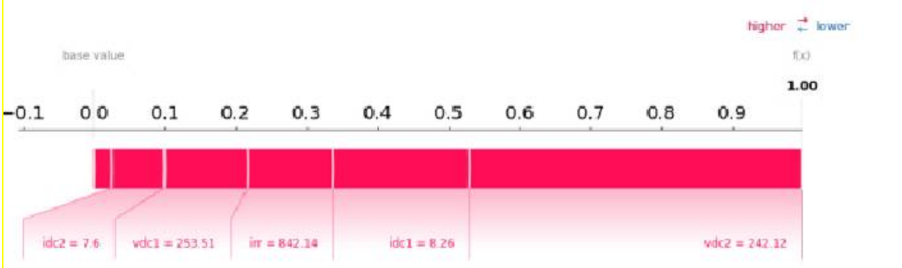
Stability evaluation results for Anchors (healthy observation).

Rule	Count
$vdc2 > 261.91$ **AND** $idc2 > 7.56$ **AND** $vdc1 > 264.04$	26
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 6.03$	12
$vdc2 > 261.91$ **AND** $idc2 > 7.56$ **AND** $pu\iota \leq 45.59$	3
$vdc2 > 261.91$ **AND** $264.04 < vdc1 \leq 272.7$ **AND** $idc2 > 6.03$	3
$261.91 < vdc2 \leq 285.71$ **AND** $idc2 > 6.03$	2
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 7.56$	2
$261.91 < vdc2 \leq 271.99$ **AND** $idc2 > 6.03$ **AND** $vdc1 \leq 272.7$	1
$vdc2 > 261.91$ **AND** $idc2 > 7.56$	1

Short-circuit observation

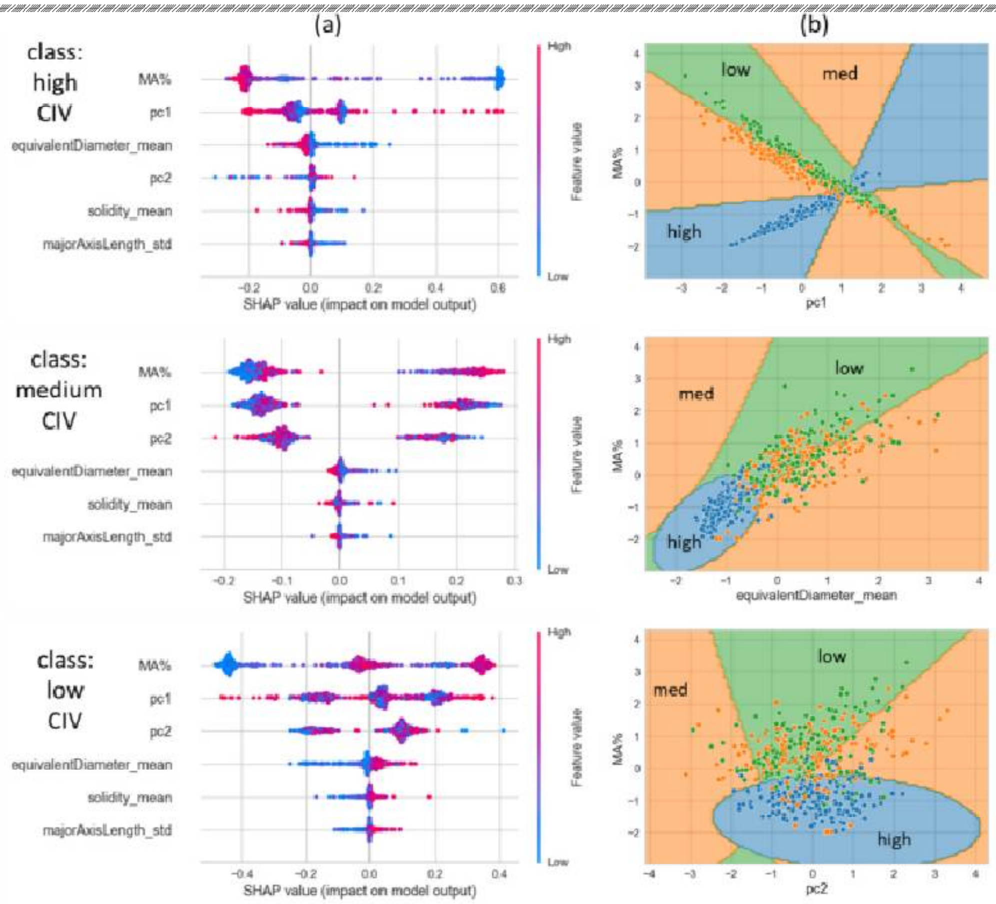


Degradation observation



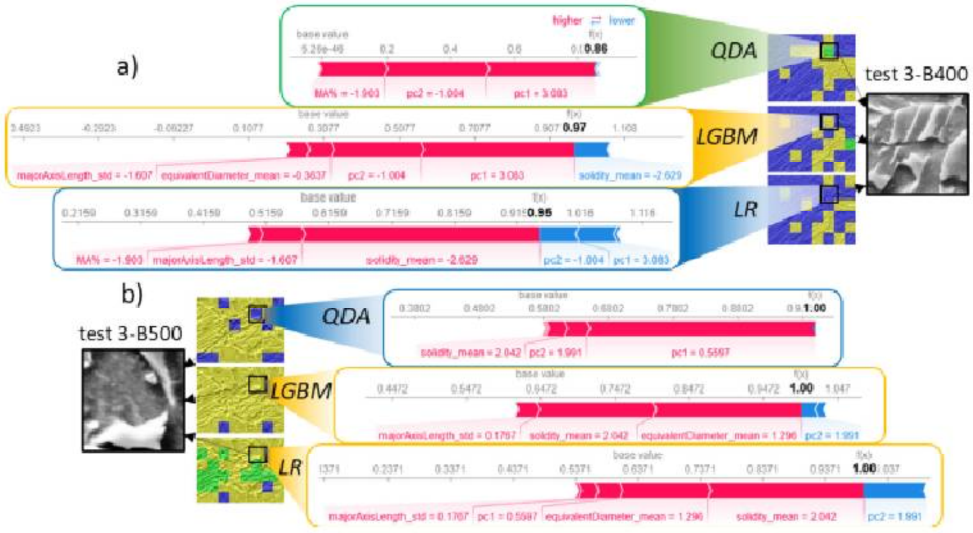
xAI.		2023-027

Global feature contributions on QDA model output

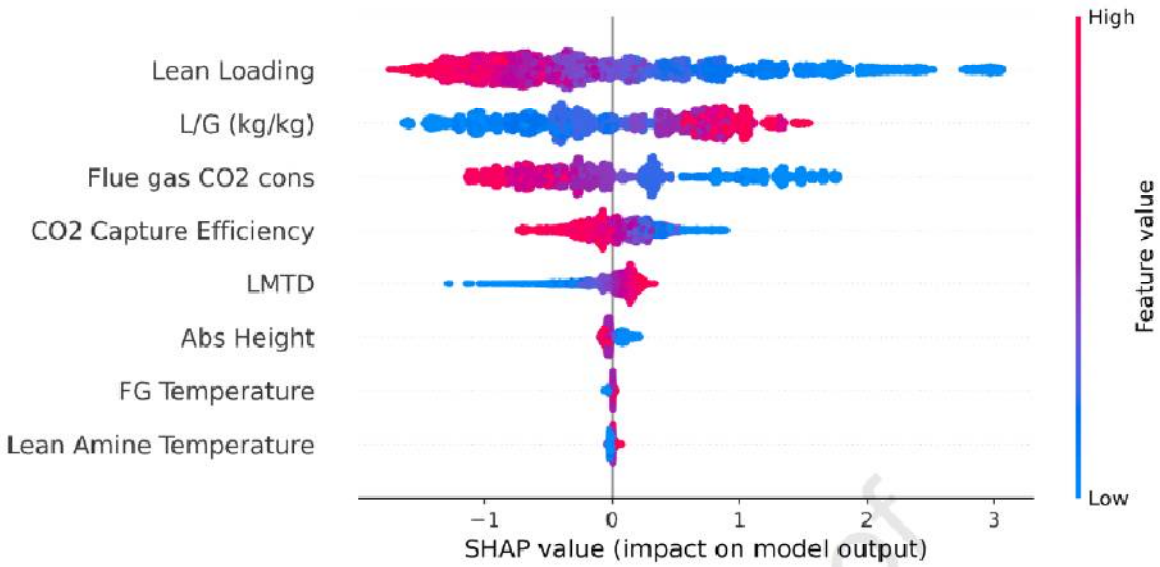


- ✓ x-axis :Feature contribution as SHAP value ;
- ✓ Colourof individual points represents the original value of the feature;
- ✓ Vertical spread over points has no meaning;

xAI.		2023-027
Local interpretability of the classified 256x256 bins for three impact energy classes		

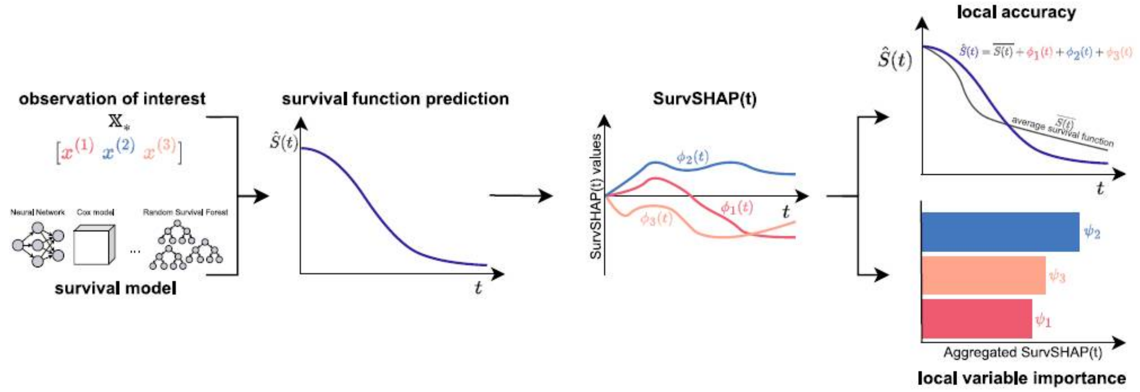


xAI.		2023-055



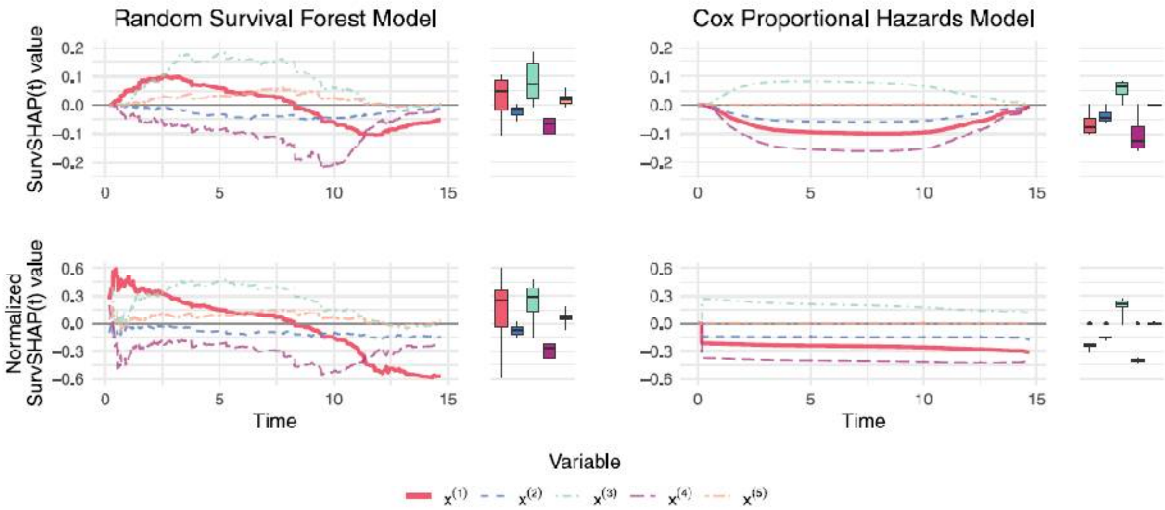
SHAP. MedData→xML(xAIM)

xAI.	Medical	2023-139
	Survival model predictions	

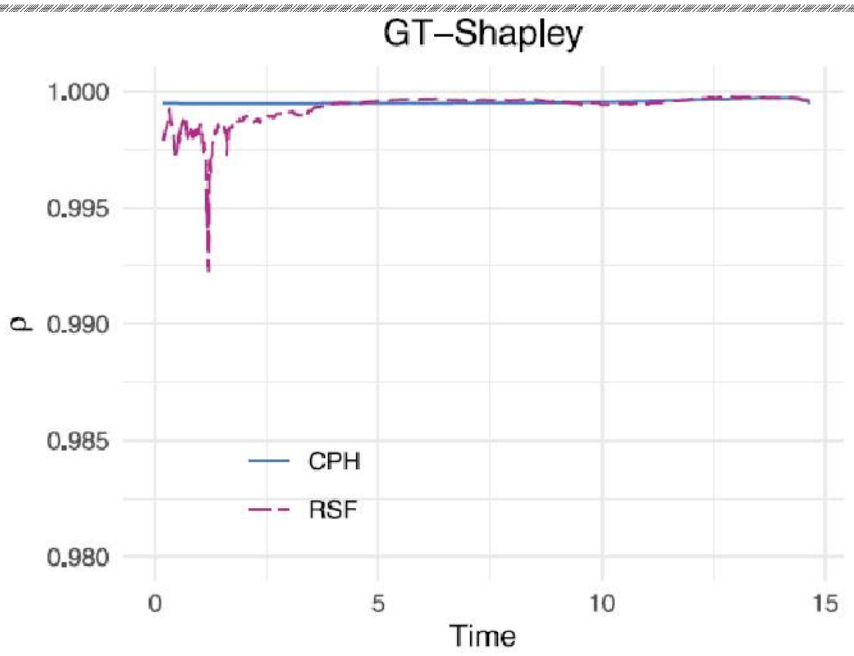


✓ SurvSHAP(t) allows for time-dependent explainability of any survival model predictions.

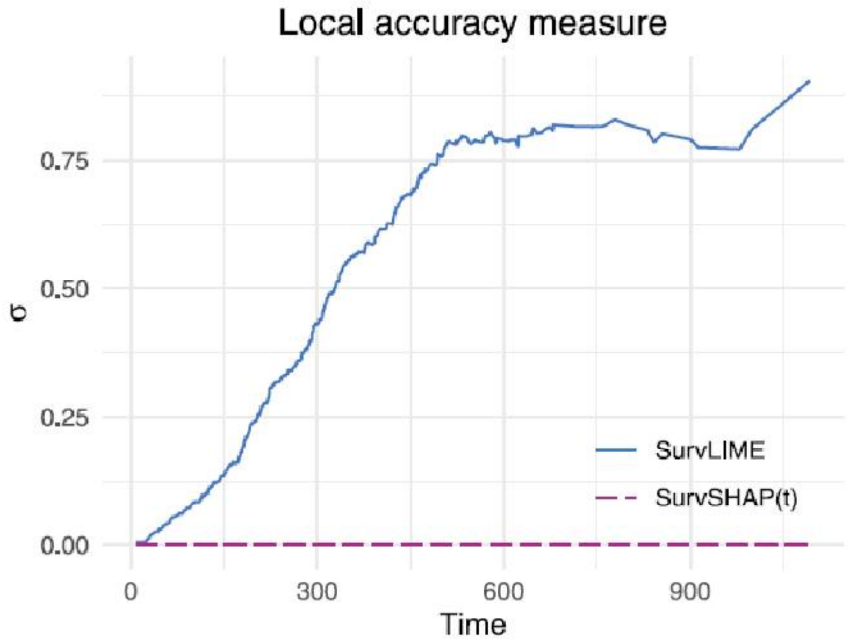
SurvSHAP(t) for the selected observation and two models



Comparison of the GT-Shapley metric for explanations of CPH and RSF trained on the EXP1_complex dataset



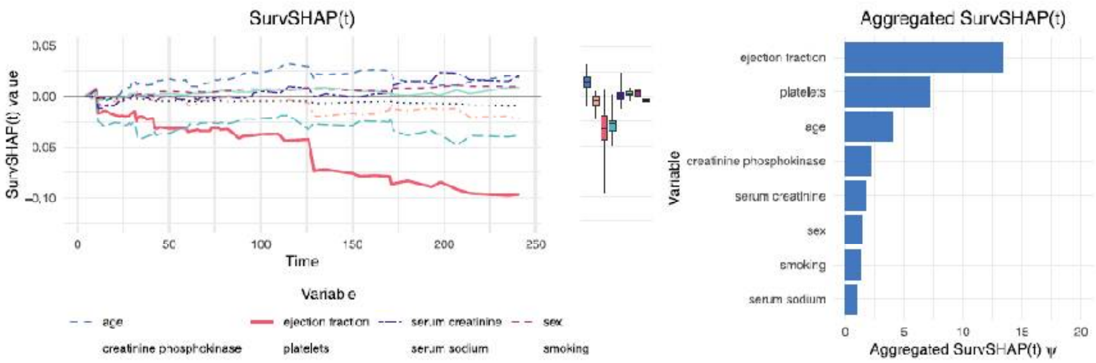
Analysis for RSF model trained on dataset0
SurvSHAP(t) trend



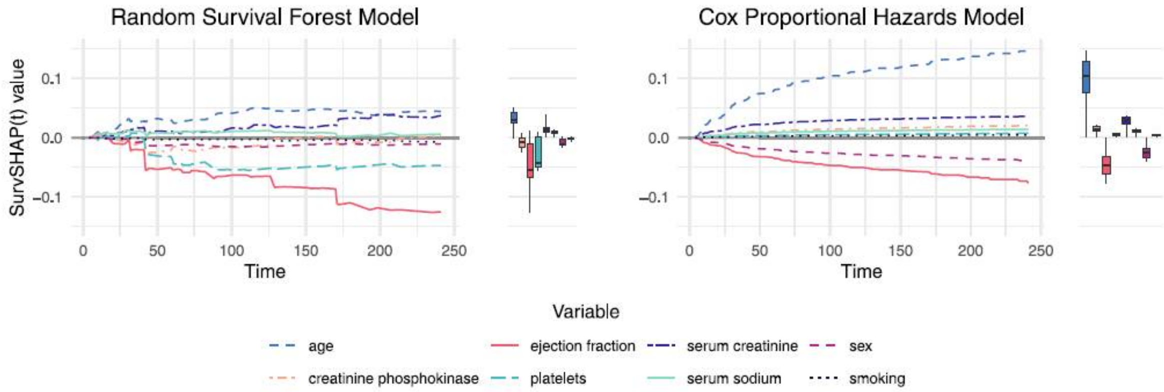
- ✓ Y axis: Normalized standard deviation of difference between black-box model output and the explanation (lower is better).
- ✓ curve for SurvSHAP(t) coincides with the x-axis

xAI.	Medical, heart_failuredataset	2023-139

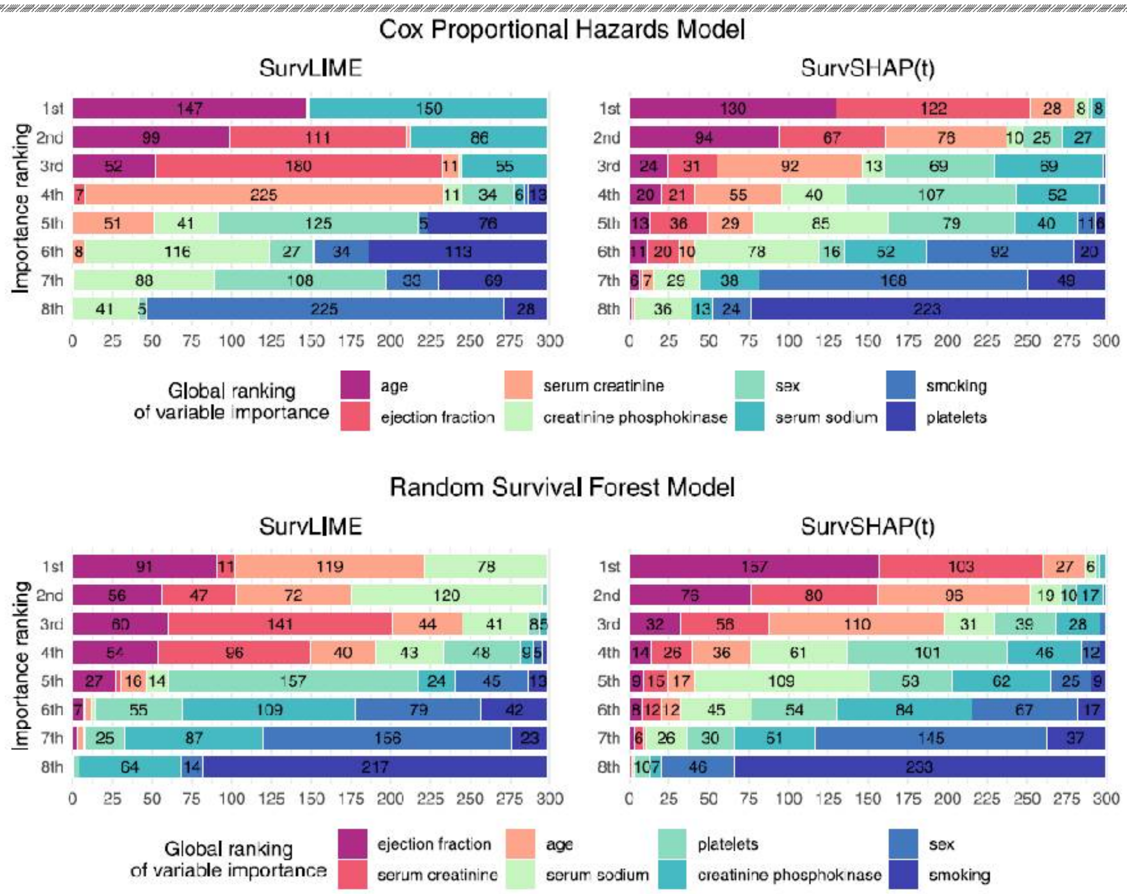
Explanation for the selected observation
RSF model trained on the heart_failure dataset



SurvSHAP(t) for the selected observation and two models
trained on the heart_failure dataset

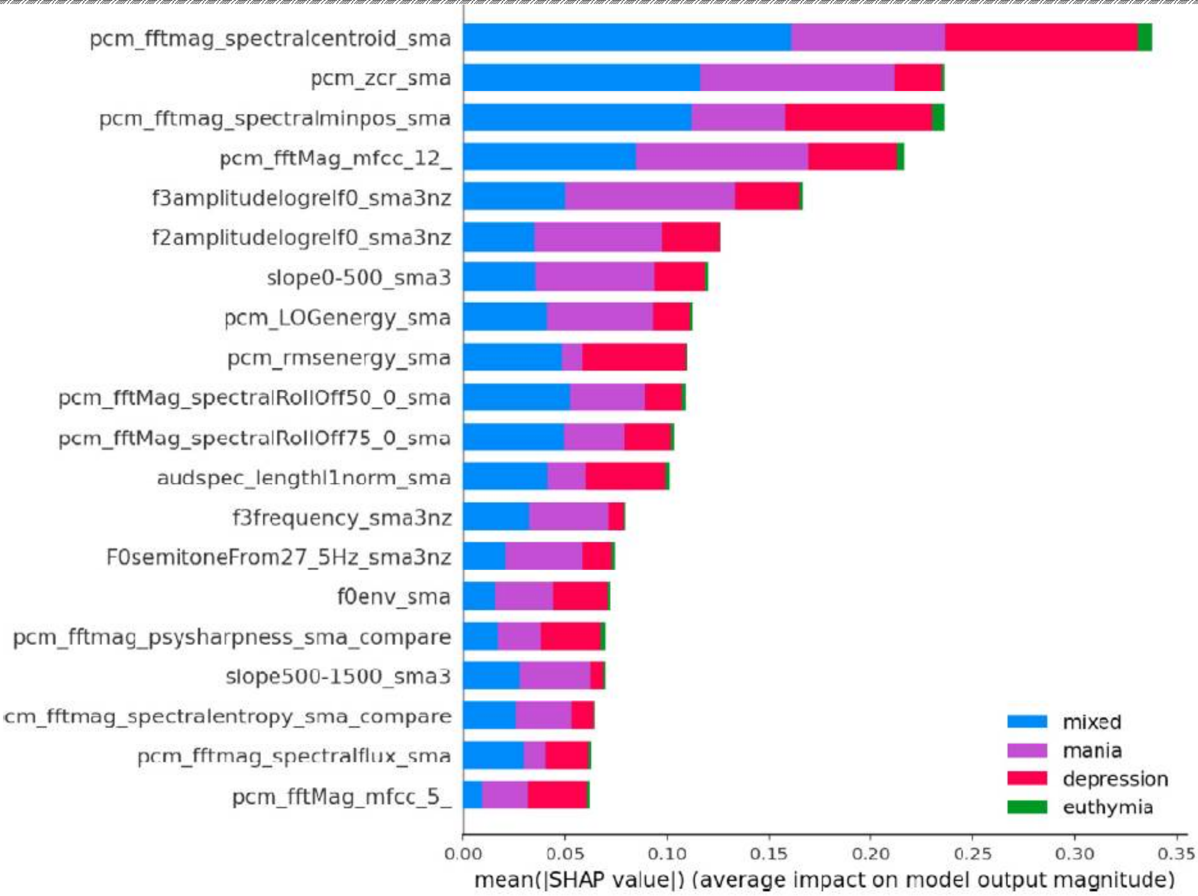


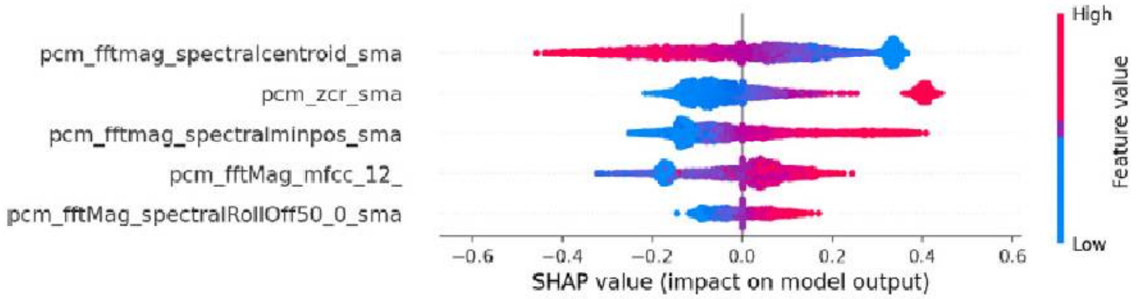
Juxtaposition of local and global importance rankings
for predictions of the CPH model (top) and RSF (bottom)

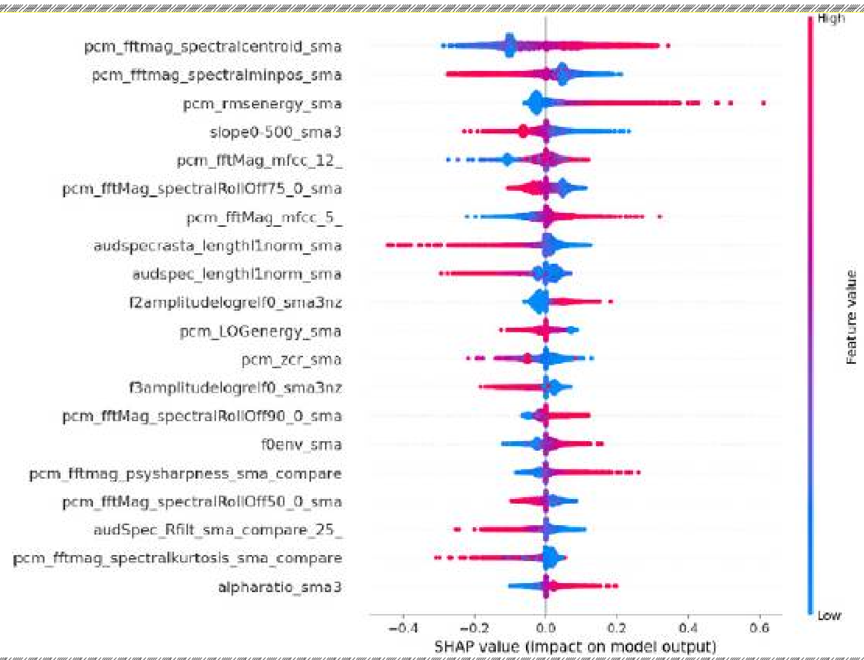


✓ Colors are specifically sorted from purpleto blue showing global ranking of variables in each model

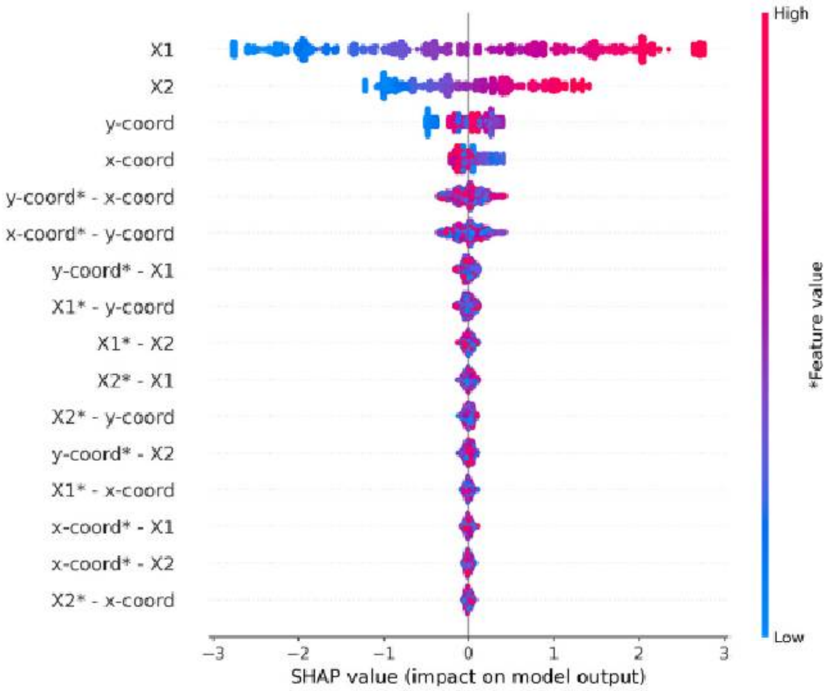
xAI.		2022-016
	Global model SHAP analysis for disease state prediction with sequential and compositional MLP model	


pcm_fftmag_spectralcentroid_sma
pcm_zcr_sma
pcm_fftmag_spectralminpos_sma
pcm_fftMag_mfcc_12_
f3amplitudelogrelf0_sma3nz
f2amplitudelogrelf0_sma3nz
slope0-500_sma3
pcm_LOGenergy_sma
pcm_rmsenergy_sma
pcm_fftMag_spectralRollOff50_0_sma
pcm_fftMag_spectralRollOff75_0_sma
audspec_lengthl1norm_sma
f3frequency_sma3nz
F0semitoneFrom27_5Hz_sma3nz
f0env_sma
pcm_fftmag_psysharpness_sma_compare
slope500-1500_sma3
cm_fftmag_spectralentropy_sma_compare
pcm_fftmag_spectralflux_sma
pcm_fftMag_mfcc_5_
mixed
mania
depression
euthymia
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mean(|SHAP value|) (average impact on model output magnitude)


pcm_fftmag_spectralcentroid_sma
pcm_zcr_sma
pcm_fftmag_spectralminpos_sma
pcm_fftMag_mfcc_12_
pcm_fftMag_spectralRollOff50_0_sma
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
SHAP value (impact on model output)
High
Feature value
Low

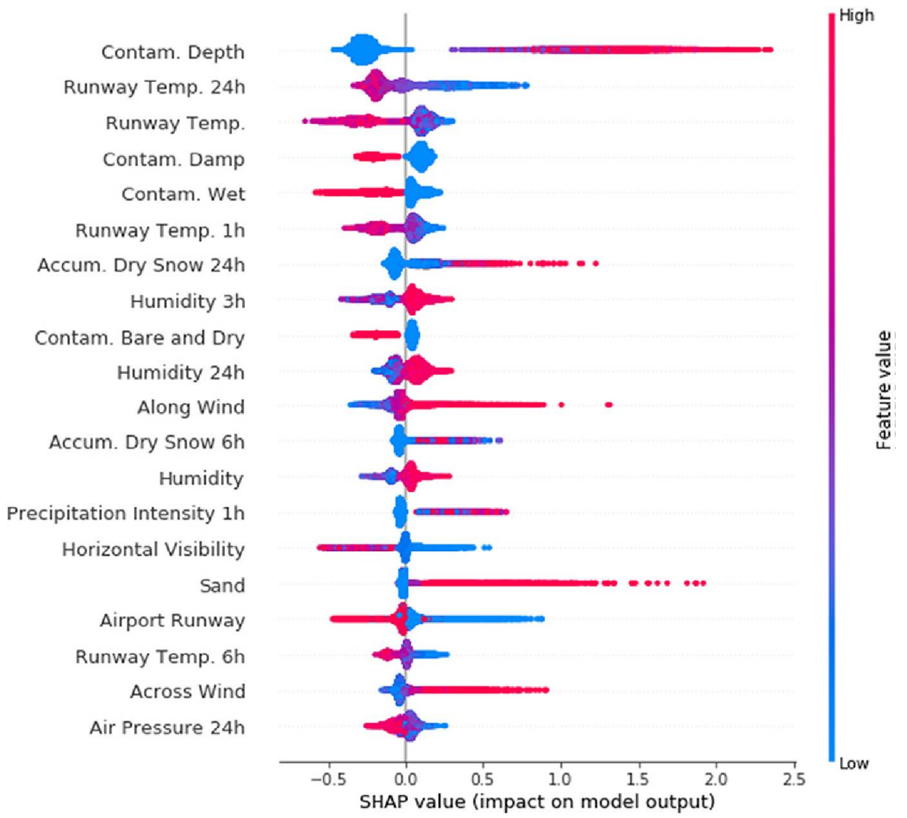


SHAP summary plot for the XGBoost model
when there is spatial autocorrelation
when there is any interaction between variables
SHAP python package

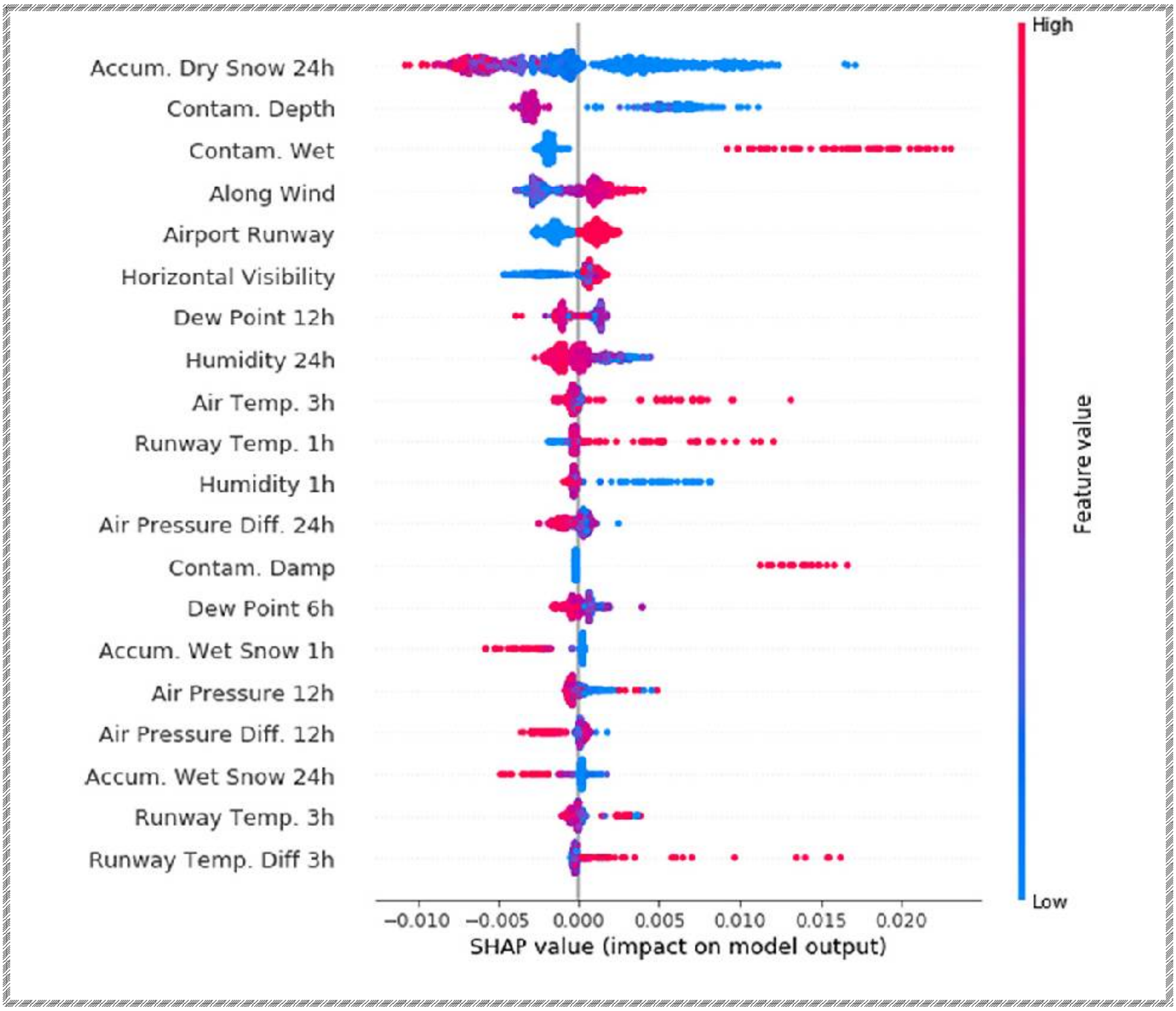


- ✓ Any interaction will appear twice and coloured by feature value noted with "*", respectively
- ✓ Ex: "X1* - X2" shows the interaction effect between feature X1 and X2 and coloured by the feature value of X1

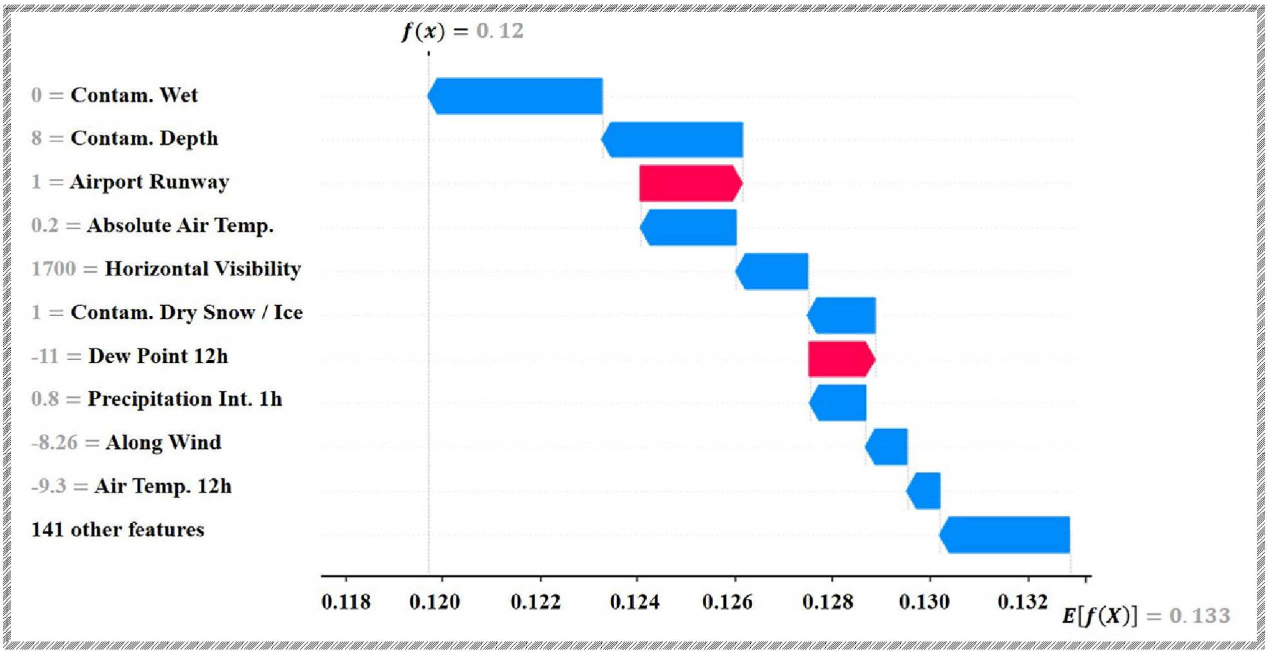
Higher SHAP values correspond to an increase in the probability for the conditions to be slippery



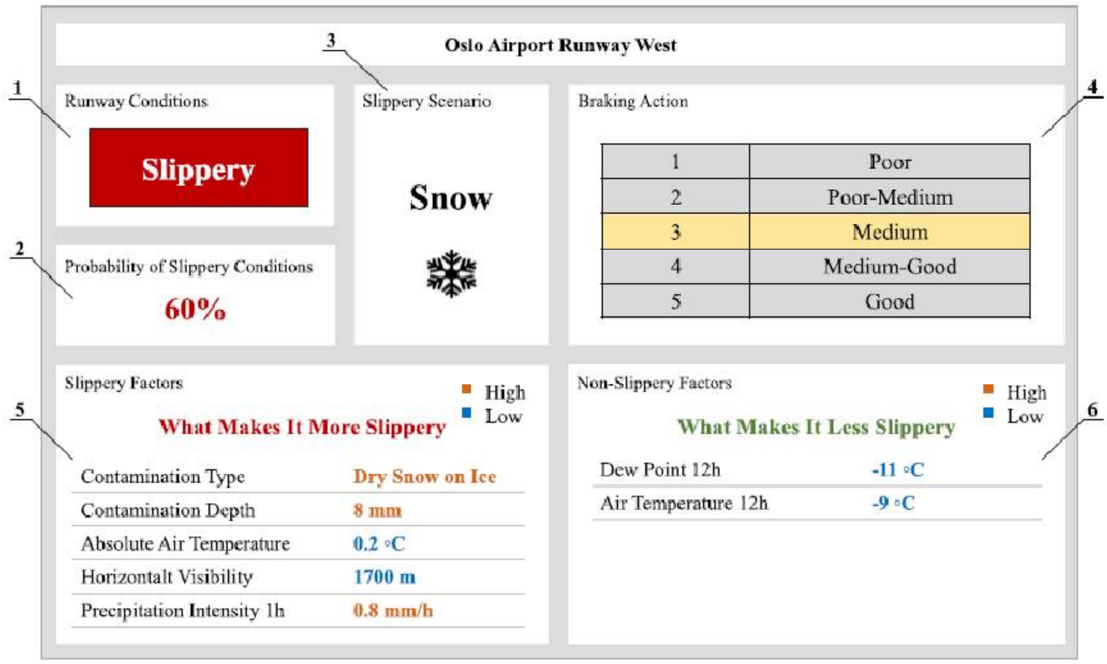
Higher SHAP values on regression model correspond
to an increase in the friction coefficient



xAI.		2022-098
	Local SHAP values for a prediction of runway surface friction at Oslo Airport	



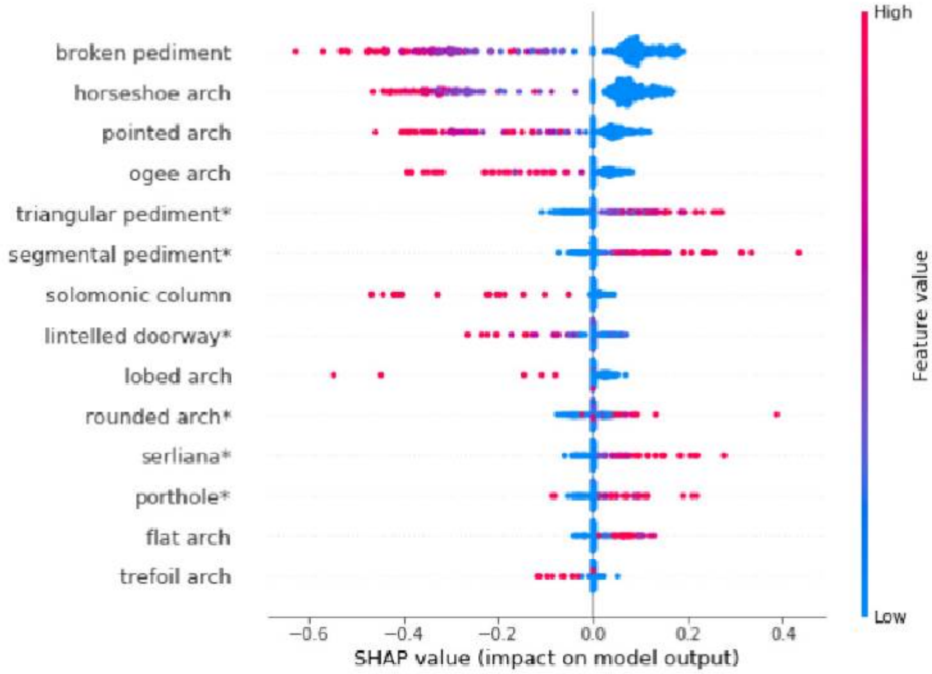
Airport runway conditions
decision support system



- Module 1 and 2 : output from classification model
- Module 3 : output from the scenario model
- Moule 4 : output from the regression model
- Module 5 and 6: output from local explanations

xAI.		2022-151

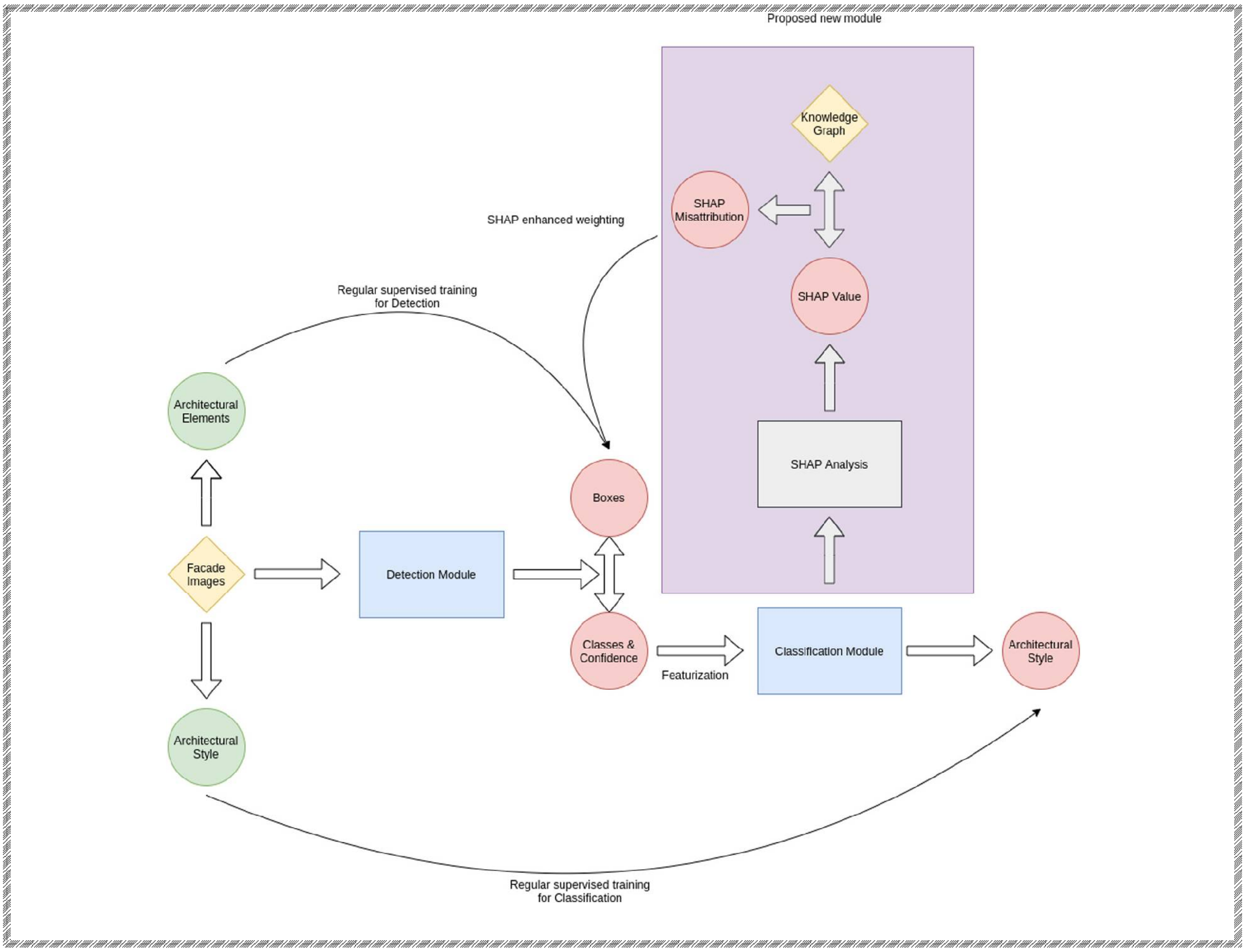
Global explanation as SHAP summary plot



xAI.		2022-151

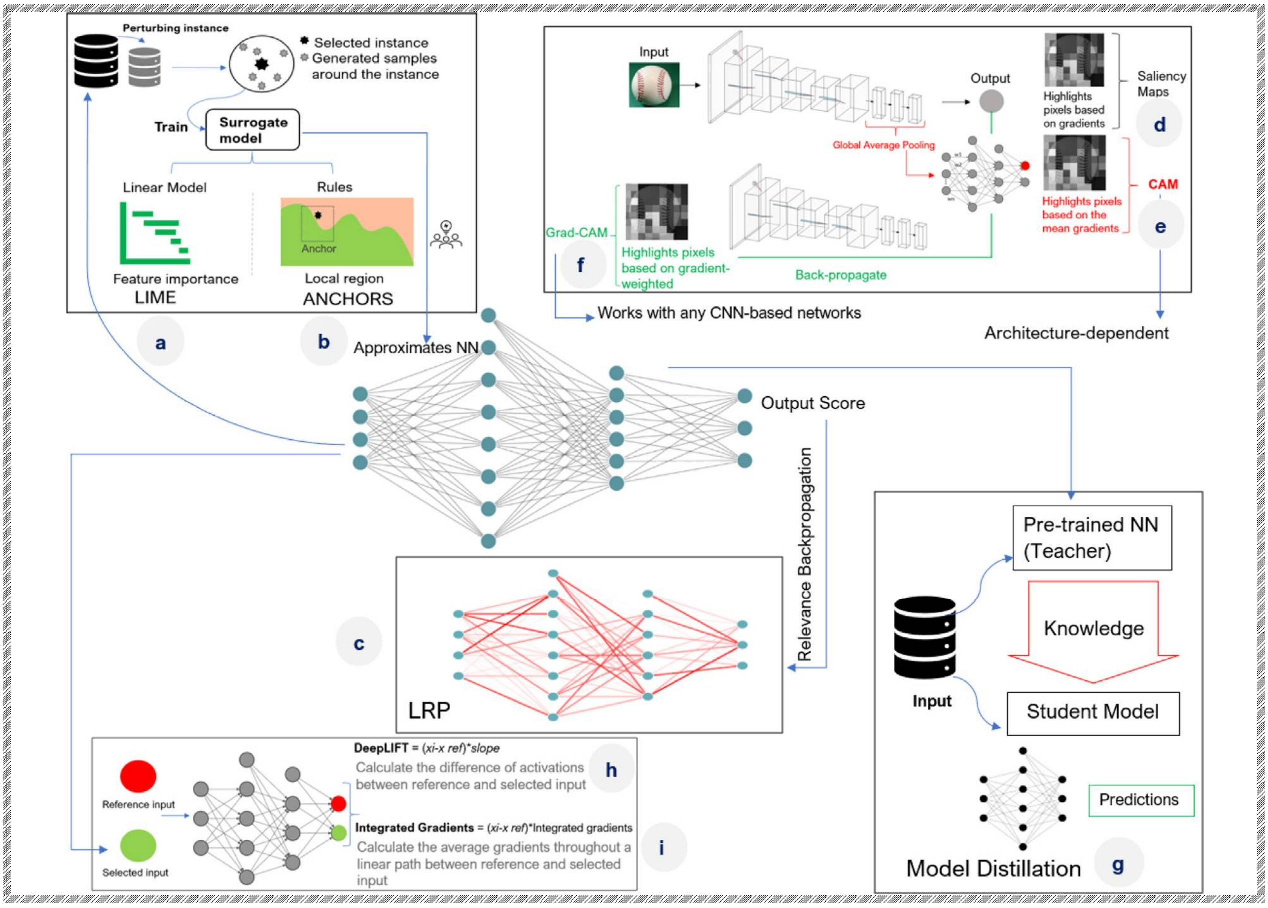
x-NeSyL(Neural-Symbolic Learning) methodology

New SHAP-Backprop training procedure



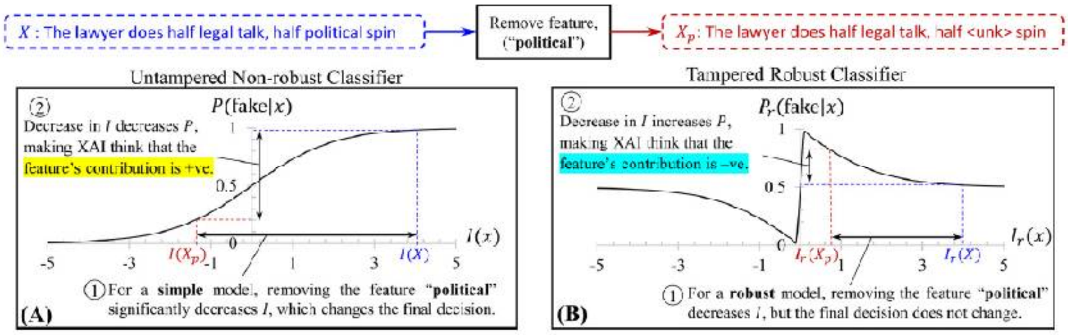
Local InterpretableModel-Agnostic Explanation (LIME)

xAI.		2022-183
	LRP-Anchors-LIME	



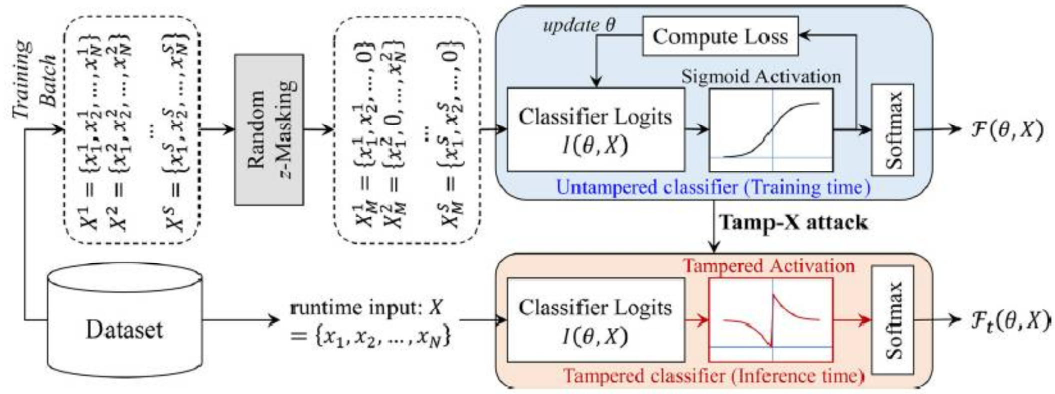
xAI.	LIME	2022-

Attack on LIME exemplified



- ✓ (A) LIME estimates the importance of input feature "political"by removing it from the input and measuring the change at the output. In this case, due to a significant decrease in the logit value, I, the classifier decision changes
- ✓ (B) When tampered activation is used in combination with the robust model, XAI method is fooled to think that the feature is contributing negatively, indicating a successful attack

Illustration of Tamp-X methodology for attacking the XAI methods



- Tamp-X first trains a robust classifier, F, using random z-masking
- Tampers the logit values, I(θ, X) using a tampered activation
- The tampered model, F t , is input to XAI method at the inference stage

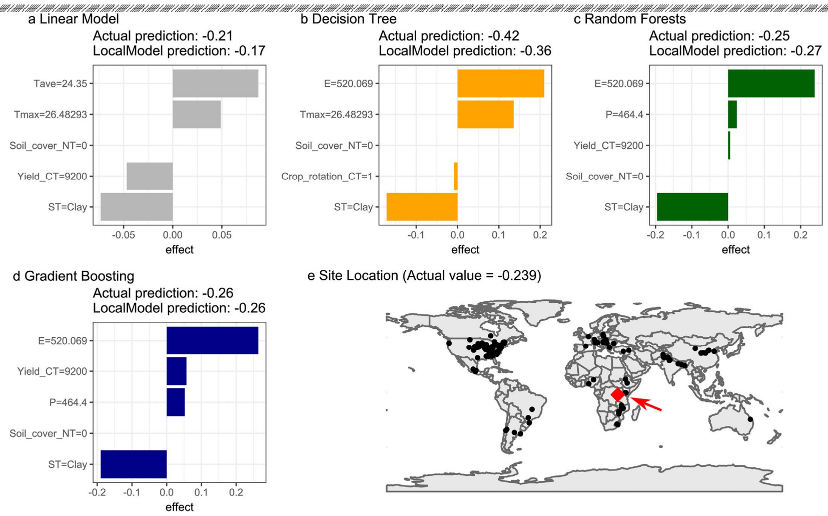
xAI.	Health care	2023-

List of XAI studies using LIME to explain model prediction results

Author	Objective	Subject	Application	Data type	Sig. features	ML/ DL	Classifier	Result
Dindorf et al. [43]	Gait Classification in Patients after Total Hip Arthroplasty	20 THA 27 normal	Gait analysis	inertial measurement unit (IMU)-based system	hip, knee, and pelvic sagittal motion and ankle rotation in the transversal plane	ML	SVM	ACC: 100%
Nanayakkara et al. [44]	Characterizing risk of in-hospital mortality	39,566 patients	ICU	EHR/EMR	lack of a motor response, low urine output, hypothermia, and higher age	ML	Ensemble	AUROC: 0.870
Uddin et al. [45]	prediction of depressive symptoms in a large textual dataset	277,552 free-text posts	Mental health	NLP	Depressed, I, Not, That, Motivation	DL	LSTM	ACC: 99.77%
Uddin et al. [46]	Human activity recognition	Public-MHEALTH 10 subjects	ADL	physiological signals (ECG + accelerometer)	Not specified	DL	LSTM	SEN: 99.00%
Magesh et al. [48]	Early Detection of Parkinson's Disease	Public-PPMI 430 PD 212 normal	CDS	Imaging data	Superpixel generation	DL	VGG16	ACC: 95.20% SPE: 90.90% SEN: 97.50% AUROC: 0.940
Palatnik de Sousa et al. [49]	Classification of Lymph Node Metastases	220,026 image patches	CDS	Imaging data	Superpixel generation	DL	VGG19	AUROC: 0.9683
Neves et al. [47]	Interpretable heartbeat classification	Public-MIT-BIH 47 subjects	CDS	1D-signal	ECG heatmap	DL	CNN	SEN: 89.50% AUROC: 0.880

xAI.	Method-- variable importance	2022-

Local Interpretable Model-Agnostic Explanations (LIME)

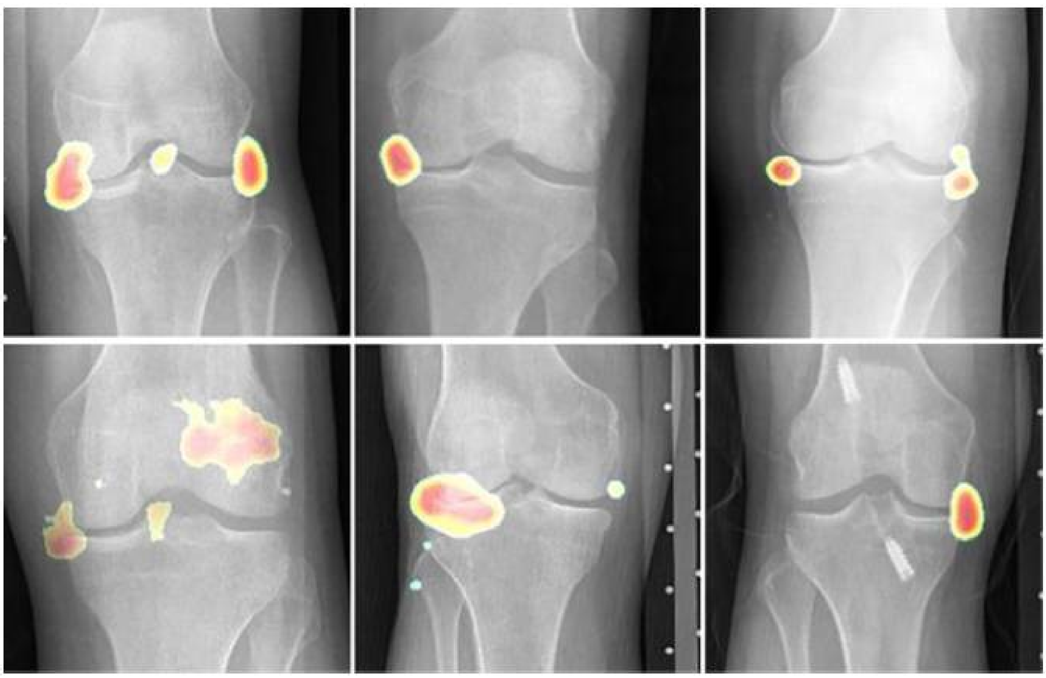


✓ LIME to explain variable importance at randomly selected local experimental site (red point in panel)

Saliency maps

xAI. | 2022-066

Knee Osteoarthritis Severity with Deep NN



Automated Classification of Radiographic images

List of XAI studies that attempted to provide explanation for NLP-based healthcare applications

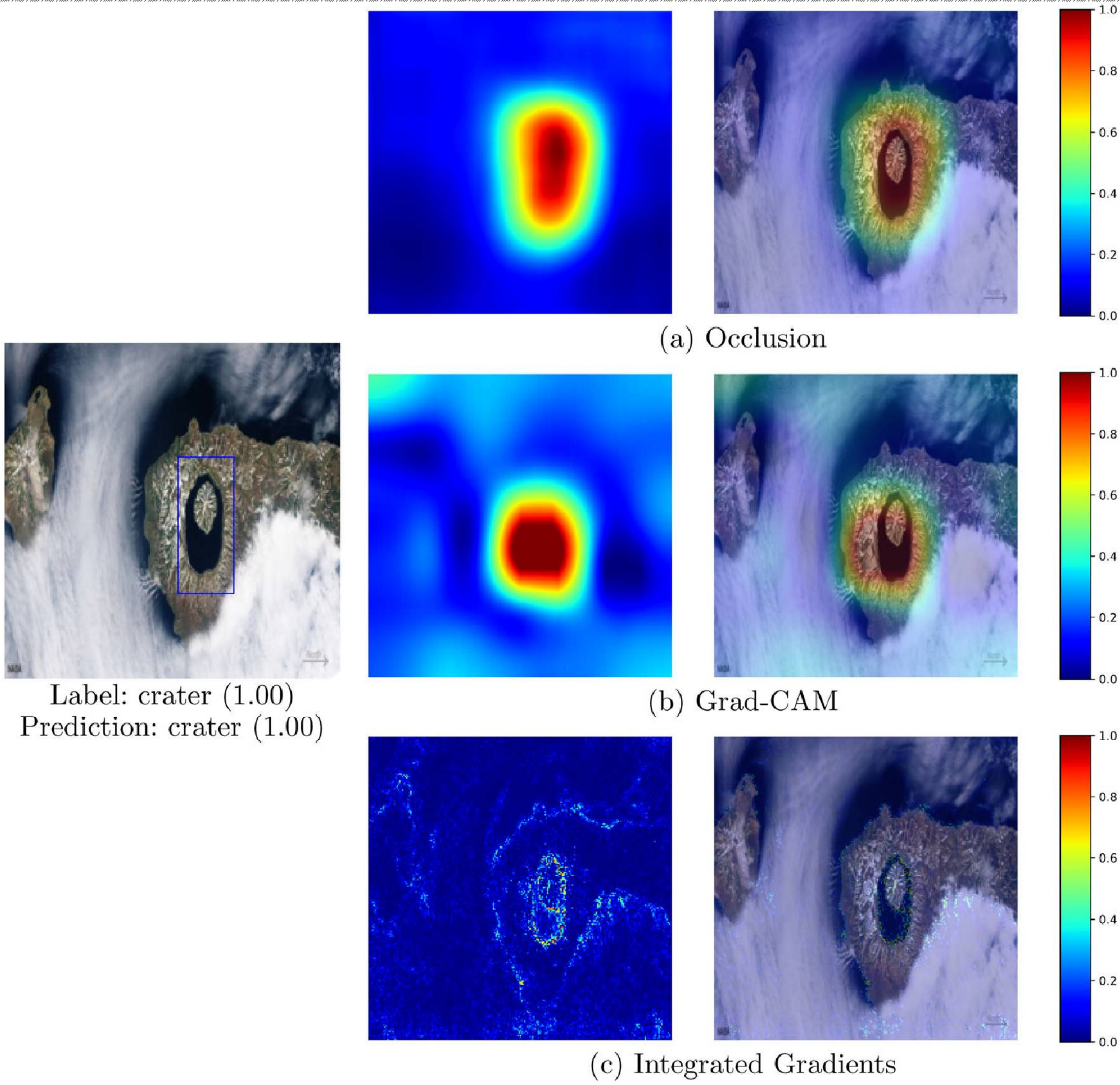
Author	Objective	Subject	Application	Data type	ML/DL	Classifier	Technique	Results
Ahmed et al. [84]	mental health treatment	Online forum, website, social media	CDS	Text	DL	BiLSTM	Attention network	SEN: 89.00% AUROC: 0.880
Dong et al. [81]	coding of clinical notes	3 Public datasets MIMIC-III, III-50, III-shielding	medical coding	Clinical notes	DL	GRU	HLAN	AUROC: 0.919
Hu et al. [83]	medical codes prediction from clinical text	Public MIMIC III 11,371 summaries	medical coding	Clinical notes	DL	CNN	Attention layer	AUROC: 0.900

List of XAI studies that used other methods to obtain saliency maps or heatmaps

Author	Objective	Subject	Application	Data type	Data	DL models	Technique	Results
Liz et al. [70]	pediatric pneumonia diagnosis	Two dataset 403 consolidation 547 non consolidation 2780 bacterial 1493 viral 1583 normal	CDS	Image	Chest x ray	CNN	Keras Vis	SEN: 72.00% AUROC: 0.890
de Souza Jr et al. [71]	evaluation of cancer in Barrett's esophagus	Two dataset 64 BE 51 early-stage esophageal adenocarcinoma	Oncology	Image	Optical coherence tomography	ResNet50	Backpropagation	ACC: 87.00% SPE: 84.40% SEN: 89.00%
DeGrave et al. [72]	COVID 19 Detection	three Datasets 3,604 patients	CDS	Image	Chest x ray	DenseNet 121	Expected gradients	AUROC: 0.995

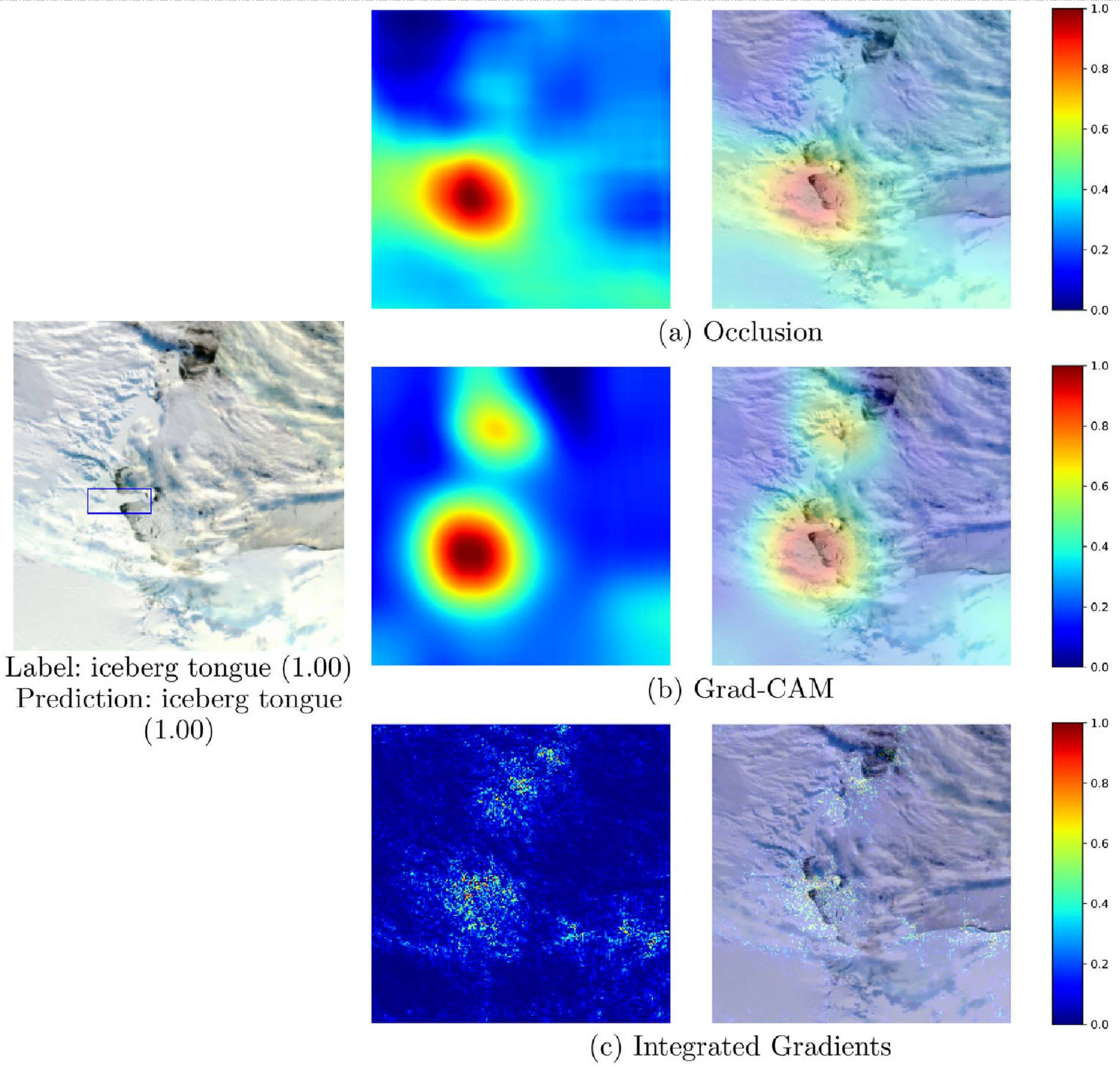
Ghorbani et al. [73]	interpretation of echocardiograms	1,624,780 video sampled images	CDS	Video	echocardiography	Inception-Resnet v1	SmoothGrad	Pacemaker AUROC: 0.890 Enlarged left atrium AUROC: 0.860 Left ventricular hypertrophy AUROC: 0.750
Chang et al. [74]	Web Diagnostic System for Phenotyping Psychiatric Disorders	288 SCZ 244 normal	CDS	Image	sMRI	DNN	z-score	Gray matter ACC: 84.00% SPE: 80.62% SEN: 89.47% White matter ACC: 90.22% SPE: 91.23% SEN: 89.21%
Gu et al. [75]	Visually Interpretable Image Diagnosis Network	888 patients	CDS	Image	CT	CNN	importance estimation network	ACC: 82.57%
Wang et al. [76]	COVID-19 Detection	Public COVIDx 266 COVID 5,538 pneumonia 8,066 normal	CDS	Image	Chest x-ray	CNN	GSI inquire	ACC: 93.30%
Gururaj et al. [77]	COVID-19 Detection	1,489 patient	CDS	Image	Chest CT	CNN	GSI inquire	ACC: 99.10% SPE: 99.90% SEN: 97.30%
Ieracitano et al. [69]	COVID-19 Detection	64 COVID 57 normal	CDS	Image	chest X-ray	CNN	Saliency	ACC: 80.00% SPE: 78.60% SEN: 82.5%

xAI.		2023-159
	Accuracy of saliency maps in highlighting shape oftarget objects	

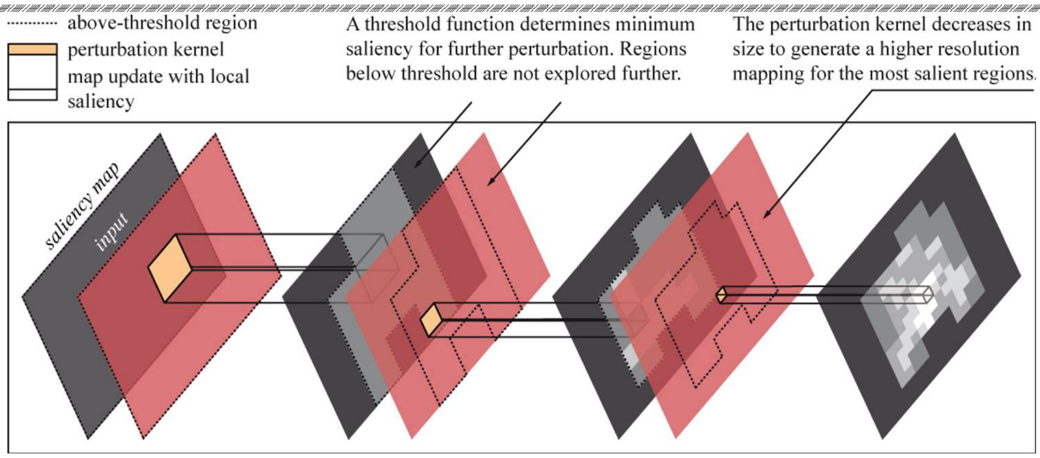


xAI.		2023-159

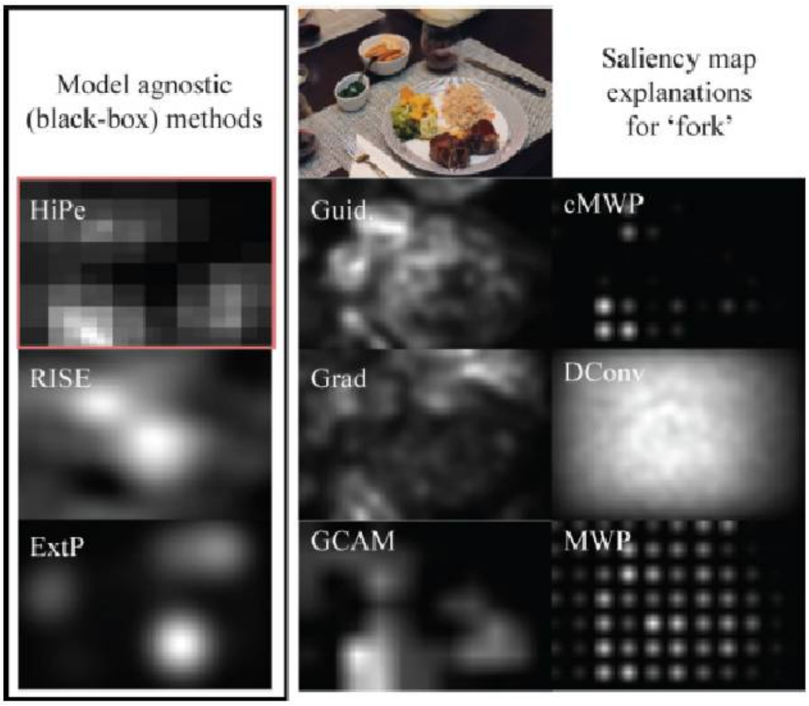
Comparison of the saliency maps in generating clear attention when
Contrast between foreground and background is low



xAI.		2022-115
Saliency Mapping with Hierarchical Perturbation		



Saliency maps for the class ‘fork’.

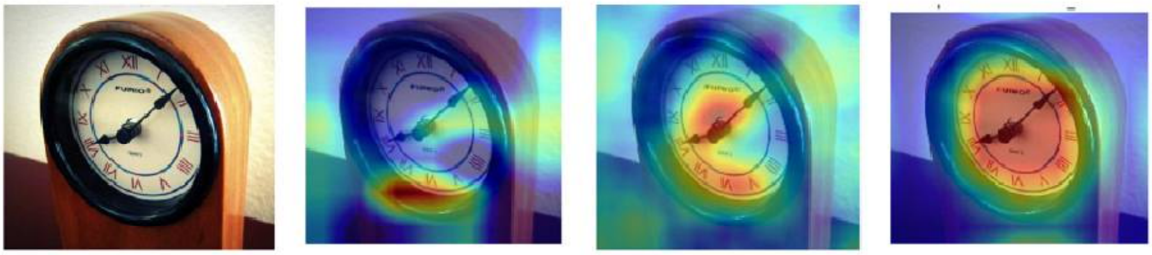


Surprising Variation In Maps Generated With Different Methods

Failure of Saliency maps

xAI.		2022-124

Example of failure of saliency maps to capture entire object class ‘clock’

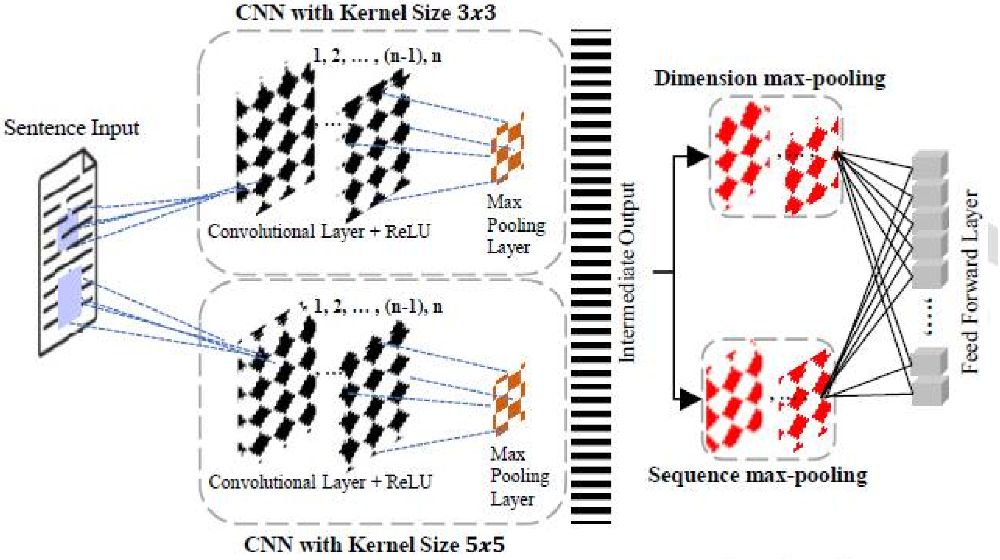


(a) Original image b) GRAD-CAM (c) RISE (d) SIDU (proposed)

Saliency Learning

xAI.		2023-030

Saliency learning



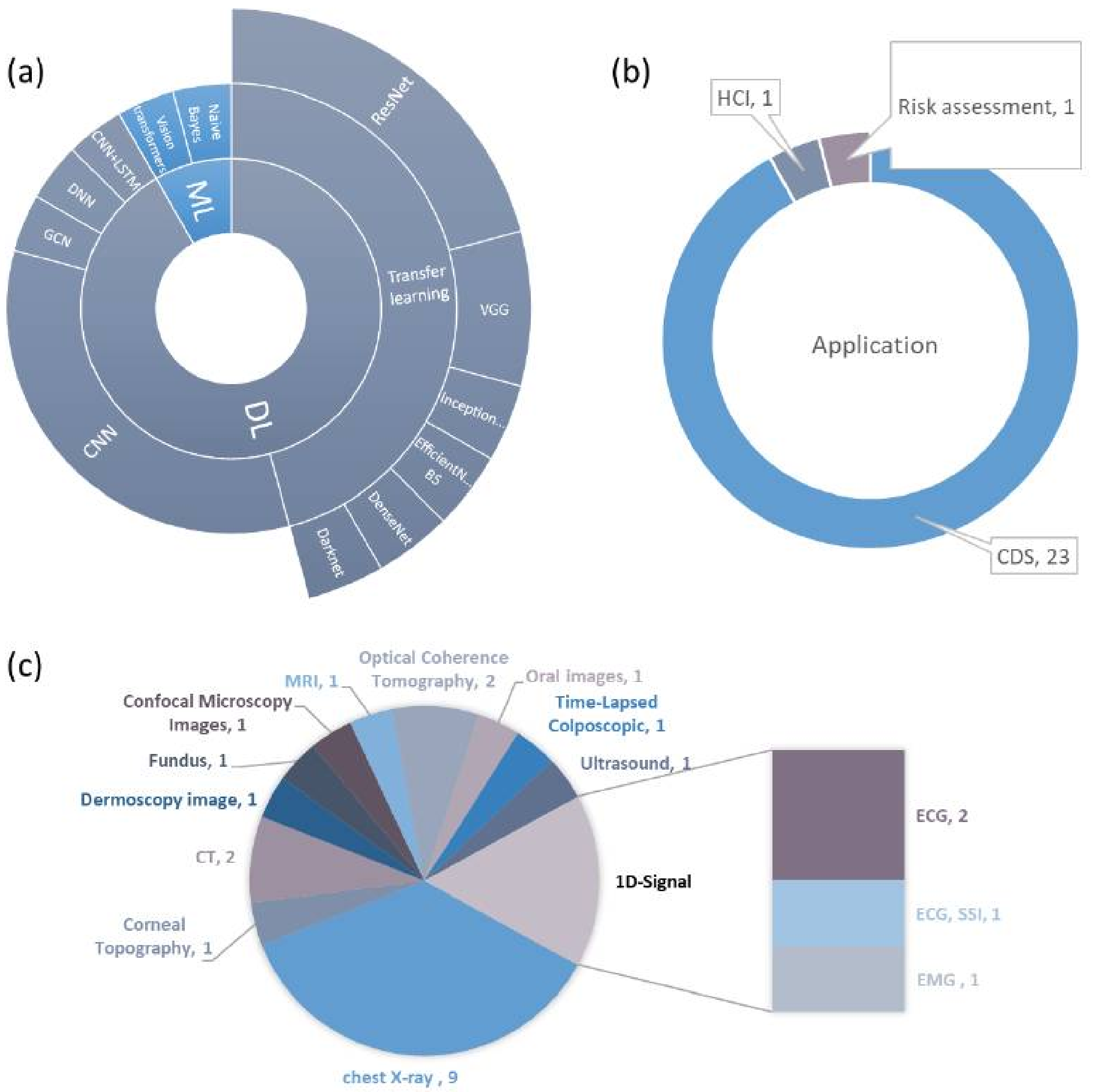
Saliency learning:

- ✓ A phrase is fed to two CNNs with kernel sizes of 3 × 3 and 5 × 5.
- ✓ Initial max-pooling operation produces an intermediate result
- ✓ Result is decomposed by performing dimensional and sequential max-pooling operations.
- ✓ Decoded output is concatenated and sent via a feed-forward layer→
- ✓ Final prediction

CAM

xAI.		2023-150

xAI models

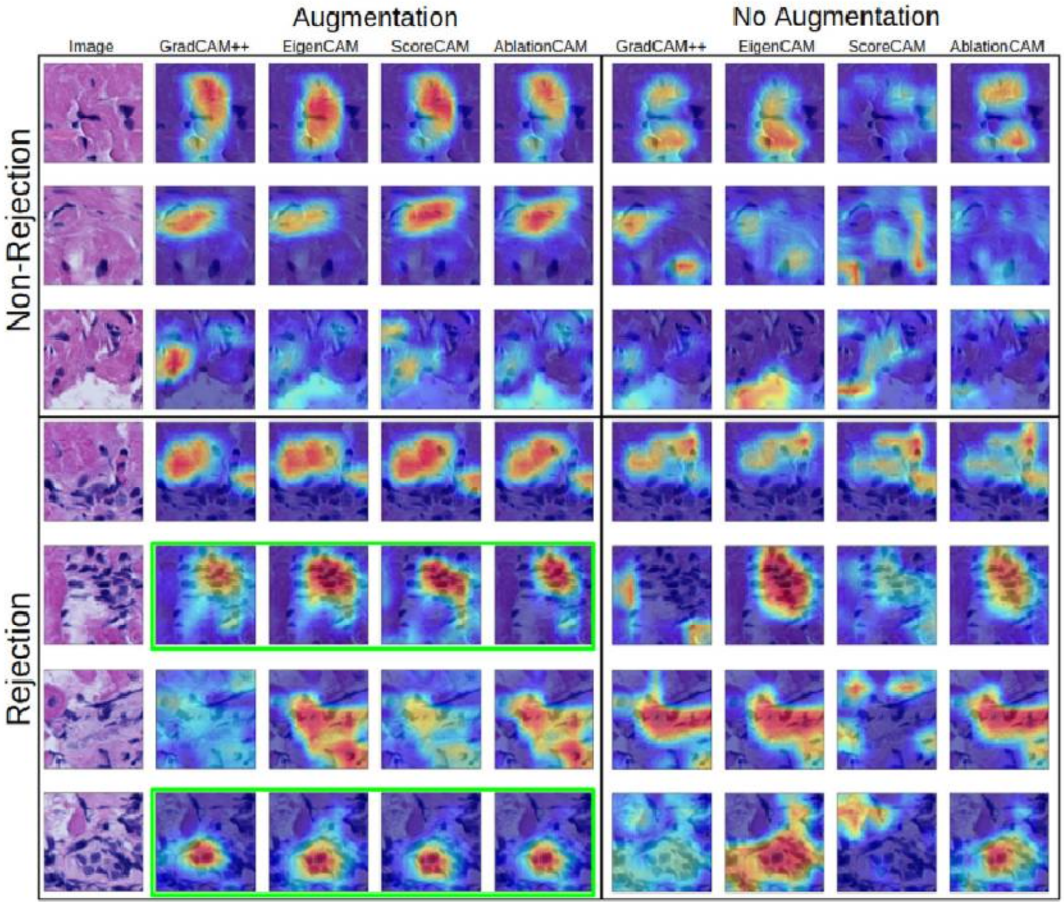


(a)Sunburst diagram of AI models used in GradCAM studies;
First level represents the type of AI model used
Second level represents the type of classifier proposed
(b) Doughnut diagram of GradCAM-studied healthcare applications
(c) A pie chart diagram illustrating the type of dataset used in GradCAM studies

xAI.		2023-151

Important regions for model labelling were highlighted using Grad-Cam++, Eigen-CAM, Score-CAM, and Ablation-CAM

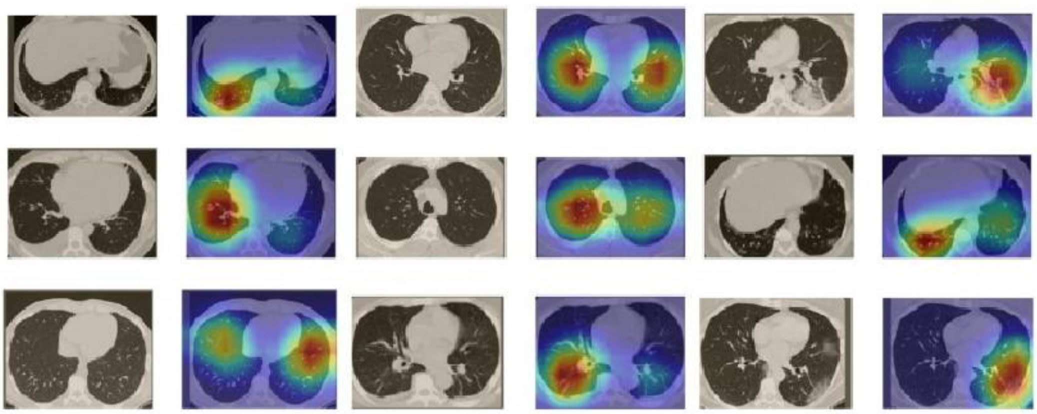
XAI methods



✓ VGG-19 models trained with and without synthetic data augmentation.

xAI.		2022-

CAM for a COVID-19 CAD model using CT imaging



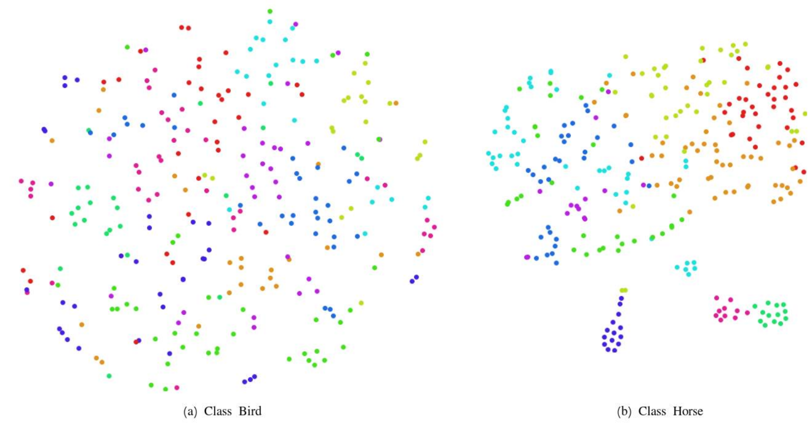
Heatmap

xAI.		2023-155

Input images of the samples in an outlier cluster of the class horse (Top)
Superimposed attribution heatmap onto Input images (Bottom)

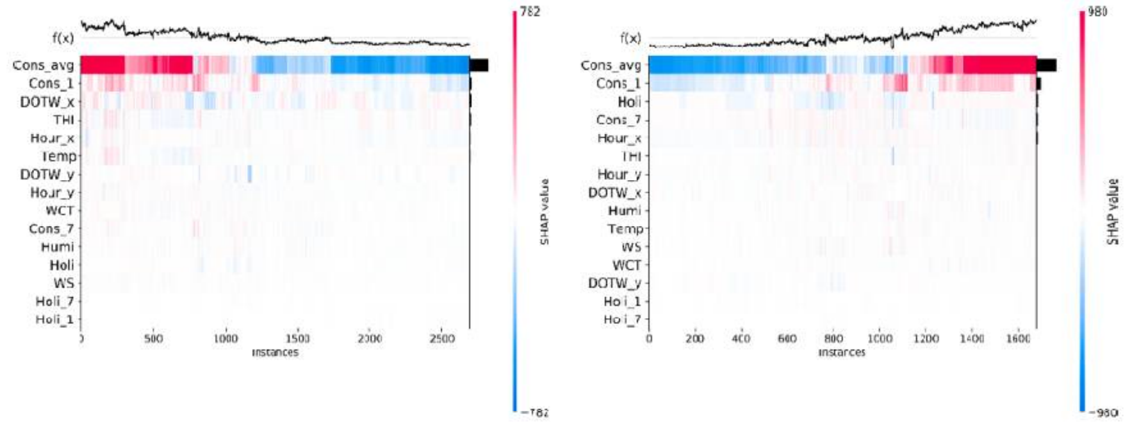
xAI.		2023-155

Comparison of t-SNE embeddings of classes bird (left) and horse (right)



xAI.		2022-

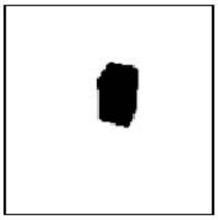
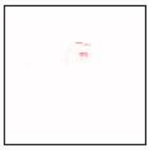
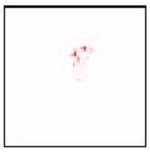
Heatmap plot
Model: LightGBM; dataset: educational building



(a) 1st Semester (b) Summer vacation

xAI.		2022-152

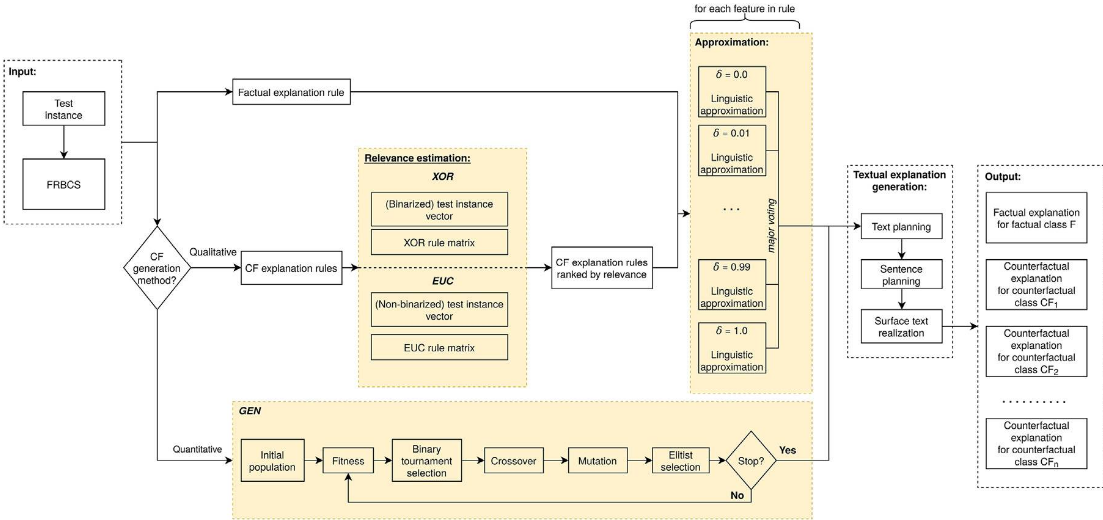
Heatmaps for a correctly predicted CLEVR-XAI-simple question (raw heatmap and heatmap overlayed with original image), and corresponding relevance *mass* accuracy.

What is the material of the large block? *metal*			GT Single Object
LRP [20]			0.94
Excitation Backprop [37]			0.82
IG [19]			0.99

Counterfactual explanations

xAI.		2023-026

CF explanation generation pipeline

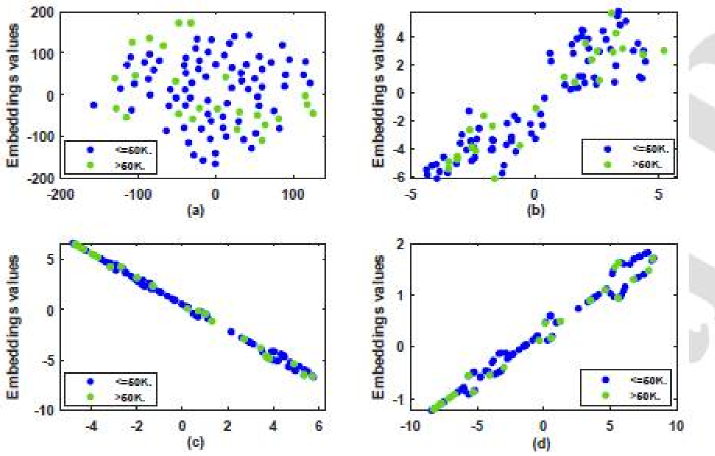


✓ Shadowed building blocks influence the surface realization of the output explanation

t-SNE

xAI.		2023-030

t-SNE



(a) Mahalanobis, b) Cosine, c) Chebychev, and d) Euclidean

- ✓ t-SNE: Produces a graph with well-defined clusters and a small number of integer data points
- + To get a better separation

xAI.	Medical	2022-

t-SNE for a COVID-19 CAD model using CT imaging