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

ISSN: 2278-1862

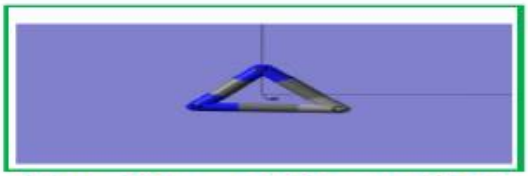
Journal of Applicable Chemistry

2023, 12 (5): 704-778

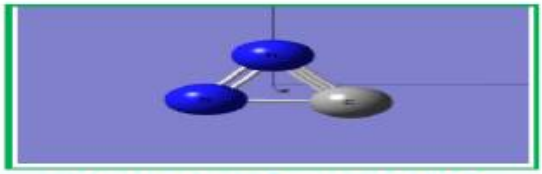
(International Peer Reviewed Journal)

New Chemistry News

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



New News of Chem (NNC)



ChemNewsNew (CNN)

CNN-56--Fit (Figure Image TableScript...)BasesPart 4. xAI (Bfit) 2022-2023 AI(1950-2023)

Information Source	sciencedirect.com;ACS.org ;	
S. Narasinga Rao M D **Associate Professor,** **Dept. of General Medicine,** **Government medical college,** **government general hospital,** **Srikakulam,** AP, **India**	**K. Somasekhara Rao, 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 start of first order logic based Symbolic Expert Systems dates back to 1960s as tools in the sub-goals of Artificial Intelligence domain. Dendral, Mycin, Xcon/Xsel were earliest the then large expert system products intended for use in organic structure elucidation in chemistry, medical diagnosis and to assist in the ordering of DEC's VAX computer systems. Xcon/Xseluse toautomatically select the computer system components based on the customer's requirements. The limitations of ESs of those days were knowledge base was human extracted and antecedents were deterministic. Further they were implemented for complex real-life scenarios. The I/O transformation was transparent and thus, explanation is straight forward to all levels of users/stake holders.

Linear models, regression trees and fuzzy-logic systems are popularly known now as machine

learning tools. Here, data flow/model structure are transparent. The vivid crystal-clear explanation renders them to be called as white-box approaches.

Neural networks of first generation viz. Adaline (Adaptive Linear Neuron), Madaline (Multiple Adaptive Linear Neuron), SLP/MLP, Fuzzy-ARTMAP, SOM, RecNN compute with floating point data and incorporating probability scores. It led to generation of robust models of high accuracy in data-driven mode. Butthe black tinge increases with complexity of model. This is the major stumble block to utilize these methods in Medicine, Defence, communication and industry.It necessiated the need to develop transparent models and explanations in multiple modes (visual, If-then-Else rules, numerical derived parameters, 2D-plots, Scripts etc.). With concerted efforts of DARPA, NSF and other agencies, a new trans-disciple called e**X**plainable Artificial Intelligence (**x**AI) emerged. The evolution of xAI is at a jet speed and serves different categories of stake holders viz. Human Experts (Hes) with specific domain knowledge, common-man using xAI-imbedded/assisted services, experts in other field but not in discipline imbedding xAI, investors, managers and policy makers involved in approvals/sanctions. The future ventures in Science/ Technology/ products or Toolswill be largely based on Trust-worthy/ Responsible/ Safe/ethical AI.

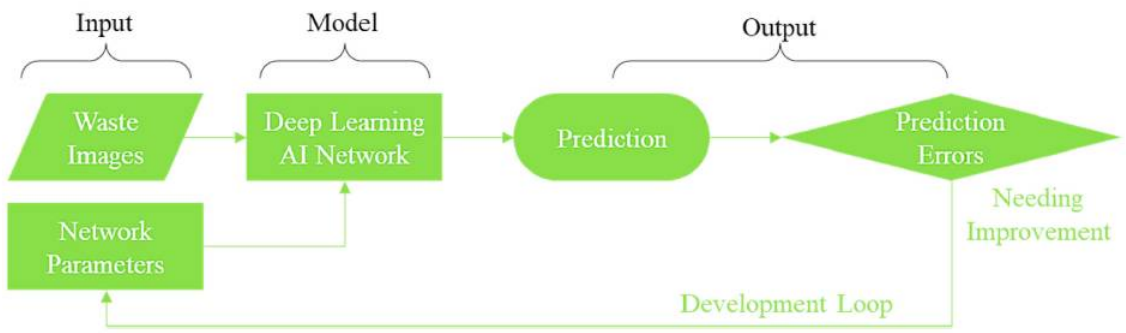
Keywords: Modelling; Artificial Intelligence; Symbolic expert systems; Second Generation AI – Neural Networks; Machine Learning; Deep Level neural architectures; Deep Learning; convolution Networks; ALEX; Capsule Nets; explainable Artificial Intelligence (xAI); Applications; DARPA Stipulations; NSF; European union;

Artificial Intelligence [1950-2023 ...]

xAI.	I(T)O.xAI Input operated by Transformer giving Output explained by xAI methods	2022-074

xAI.		2023-054

Triadic structure of AI networks: input, model and output

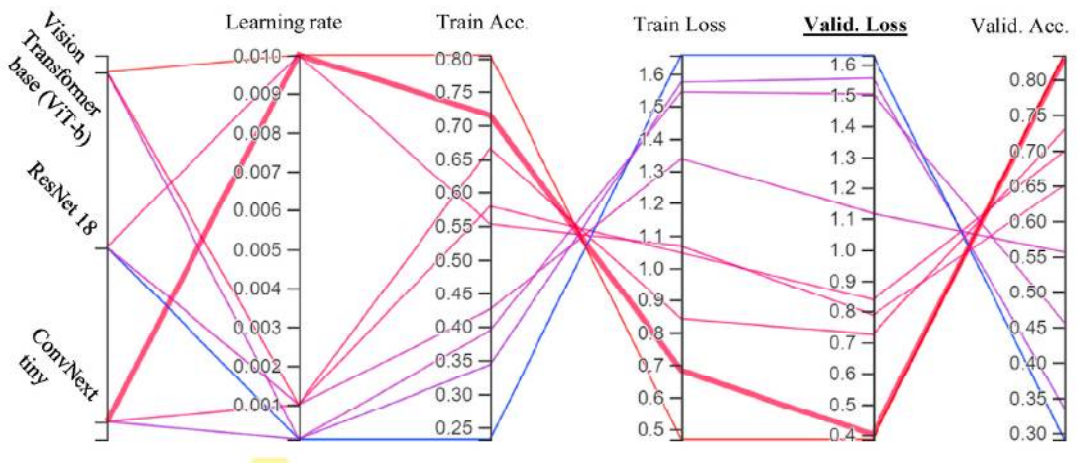


Explanation

Input, output, parameters, Network flow

xAI.		2023-054

Explaining network parameter impact on prediction performance

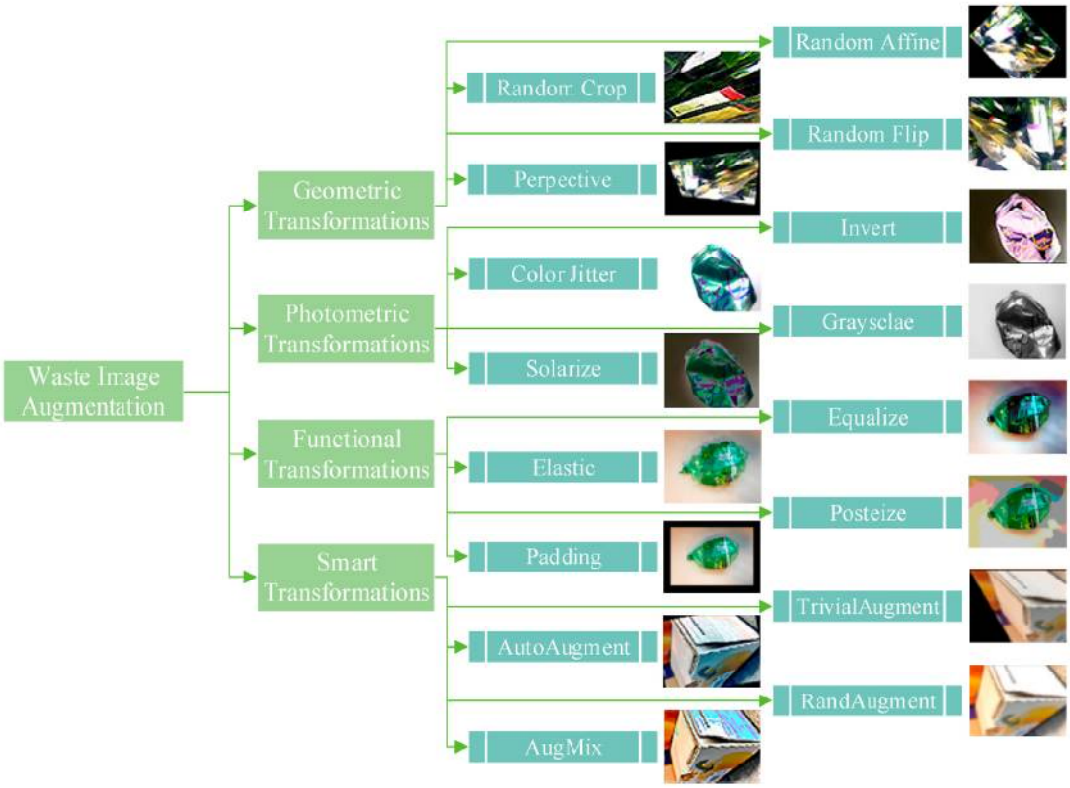


Input & Output

Augmentation.Input

xAI.	Input augmentation	2023-054

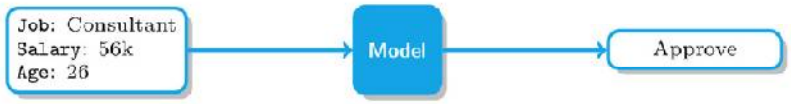
Input augmentations to maximize AI network generalizability and minimize overfitting



xAI.		2023-038

Different kinds of explanations to support human understanding

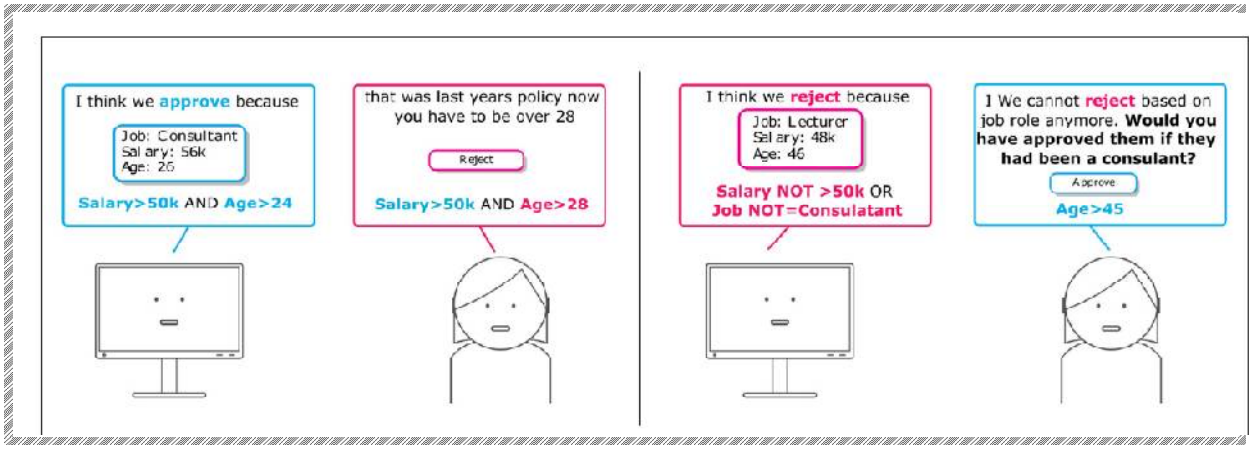
Loan approvals



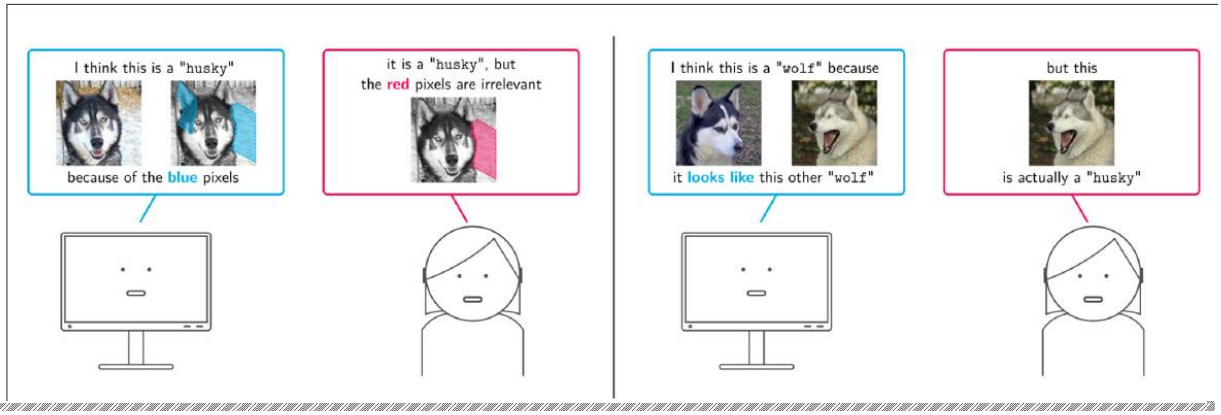
Input Attribution	Example Attribution	Rule	Concept Attribution
Importance of Job: +0.2 Salary: +0.6 Age: −0.3	Similarity to Job: Consultant Salary: 53k Age: 30 Approve	Approve if (Salary > 50k AND Age > 24) OR (Age > 45 AND Job = Consultant)	Importance of High Earner: +0.6 Professional: +0.9 Sporty: −0.2 ...

xAI.		2023-037

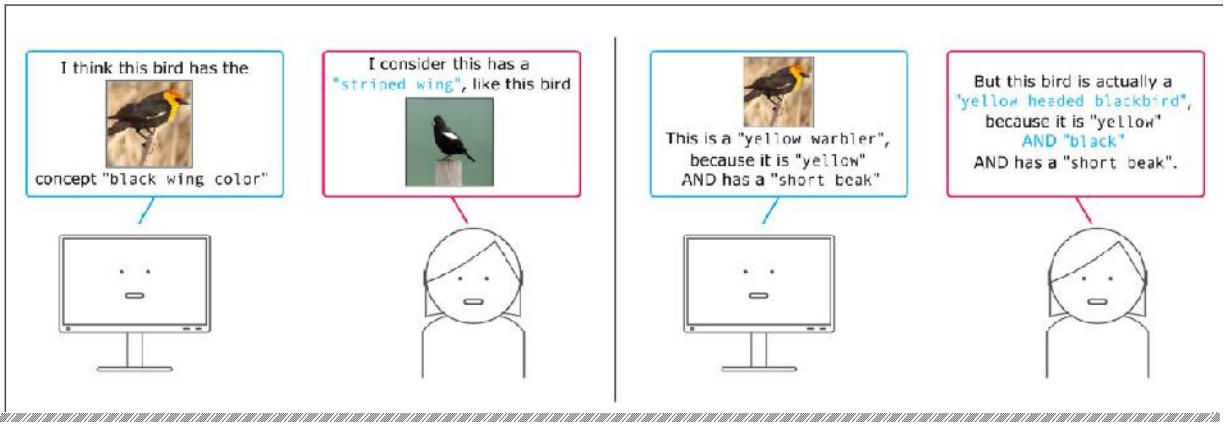
Rule-based explanations



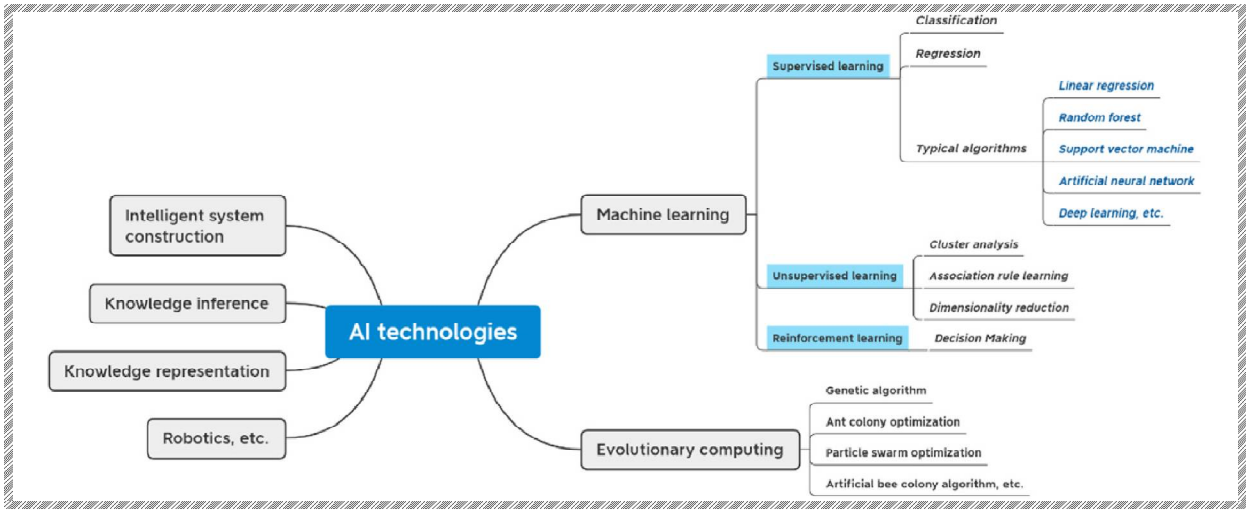
xAI.		2023-038
Predictions by highlighting relevant input Variables	Machine justifies its predictions in terms of Training examples	



xAI.		2023-038
Concept learning a model	Known basic concepts a model	

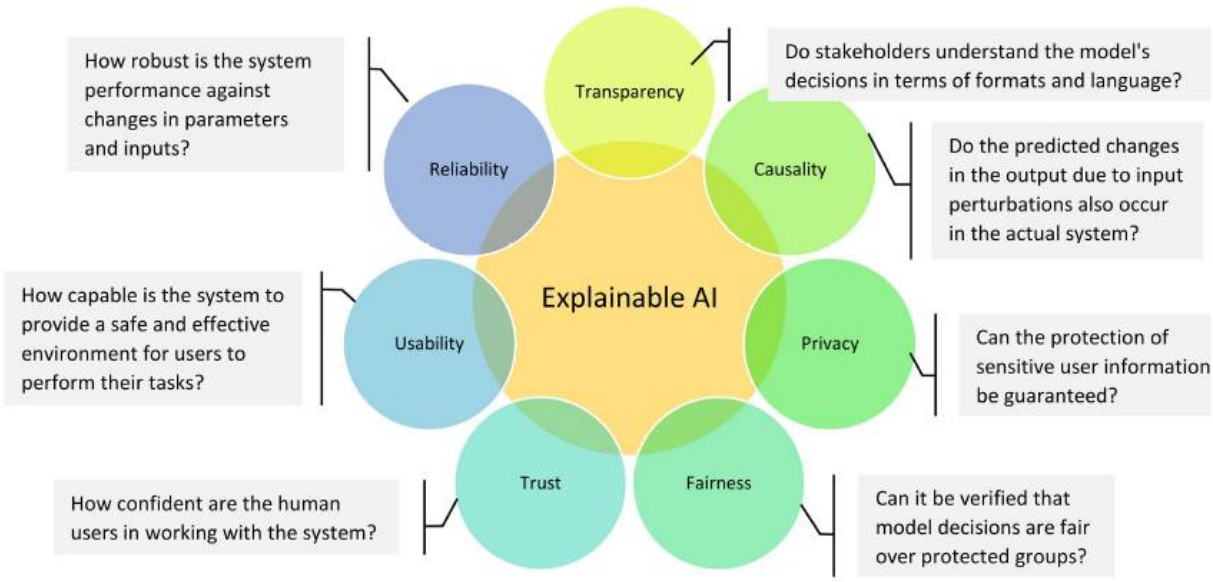


xAI.		2023-039
Framework of AI technologies		



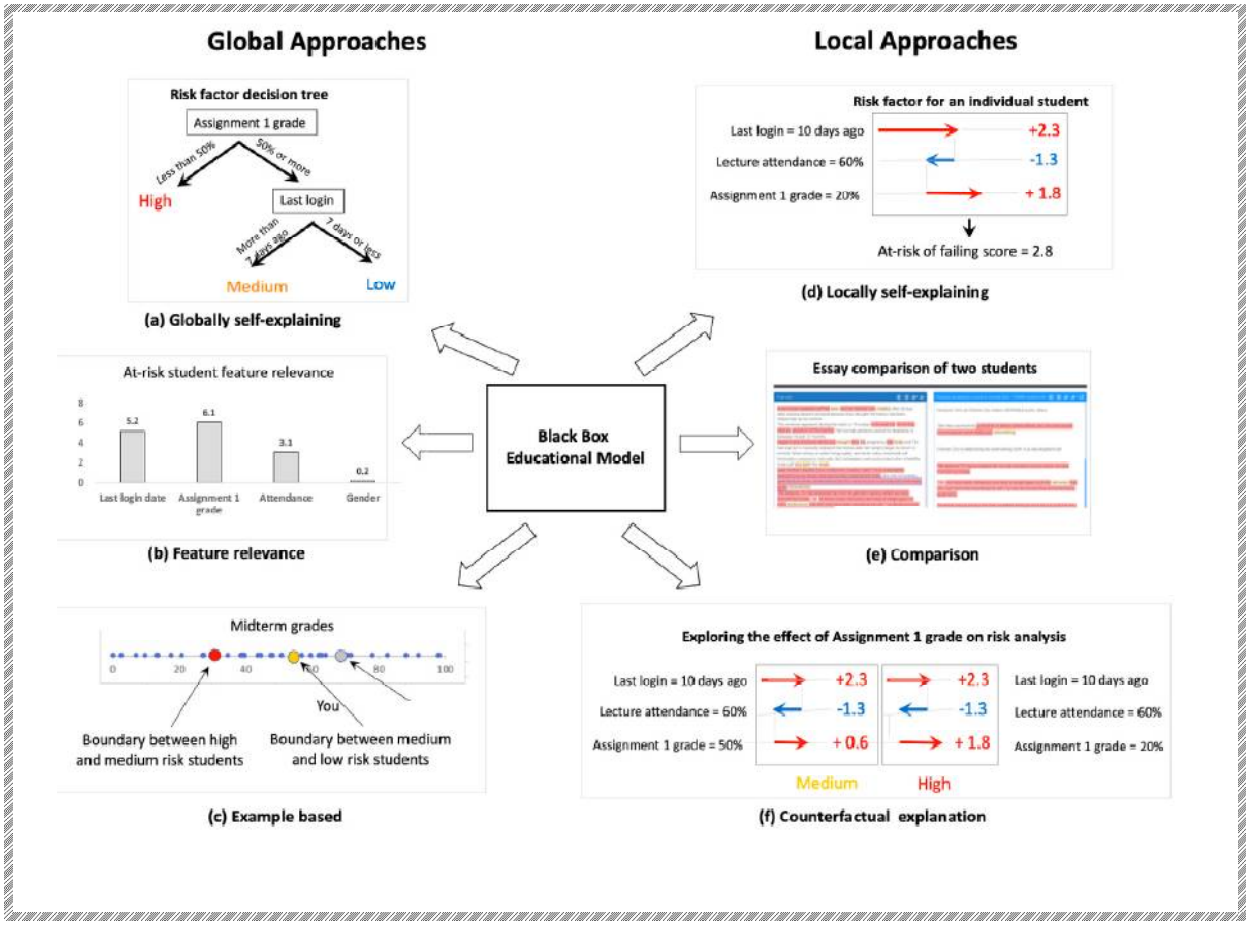
xAI. 2022-103

Requirement for interpretability in medical intelligent systems.



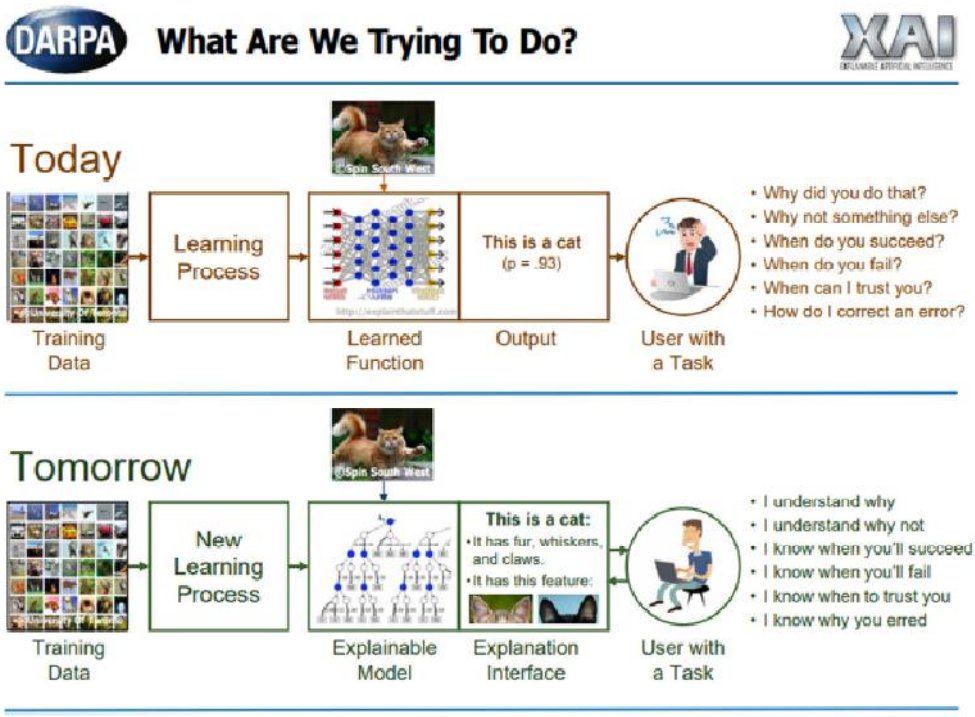
xAI. 2022-172

Explainability approaches



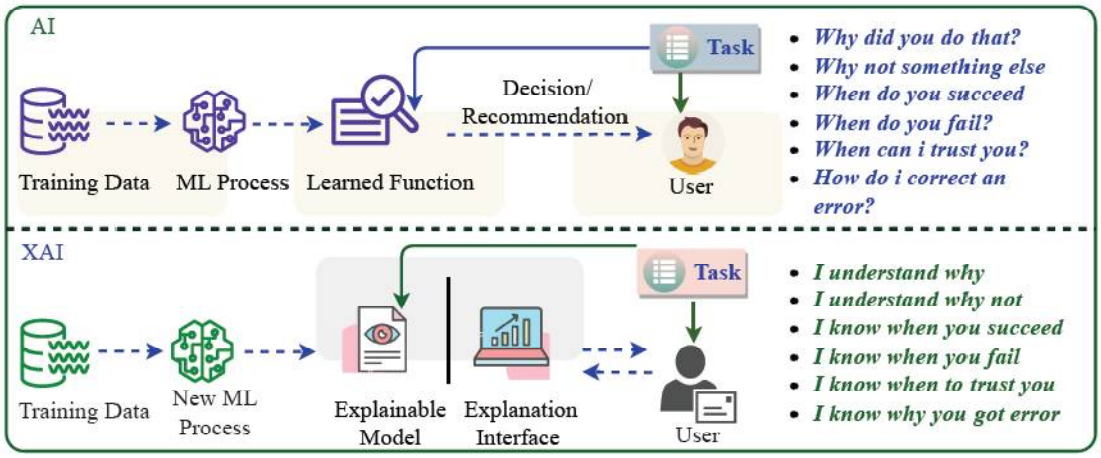
DARPA

xAI.		2022-159
Expected effect through application of XAI		



xAI.	AI vs xAI	2023-141

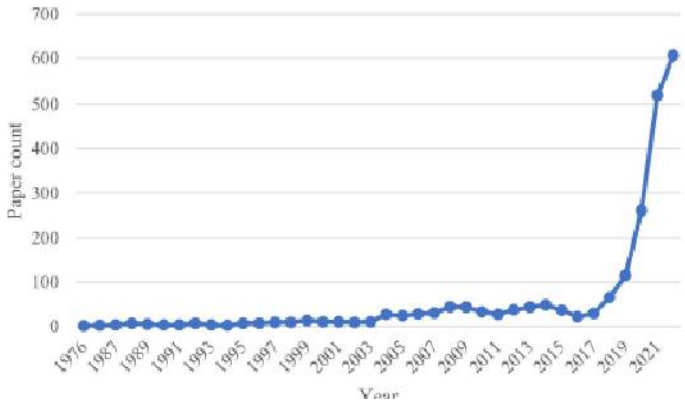
Differences in methodology of AI and XAI



xAI LiteratureSearch

xAI.		2023-105

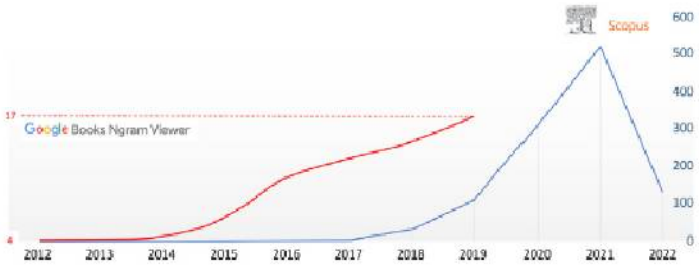
Number of XAI publications added per year from 1976 to 2021



- ✓ Database: PubMed (accessed on 1 July 2022: https://pubmed.ncbi.nlm.nih.gov)
- ✓ Terms searched: (explainable AI OR explainable artificial intelligence) AND (medicine OR healthcare)

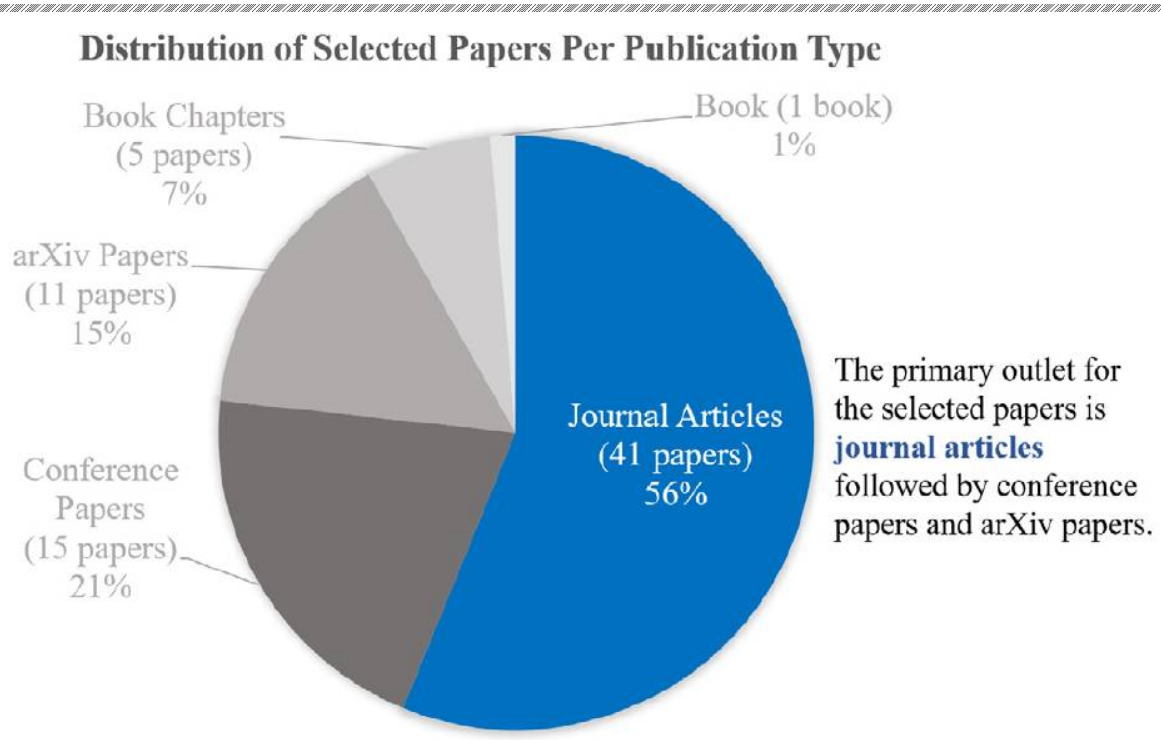
xAI.		2023-116

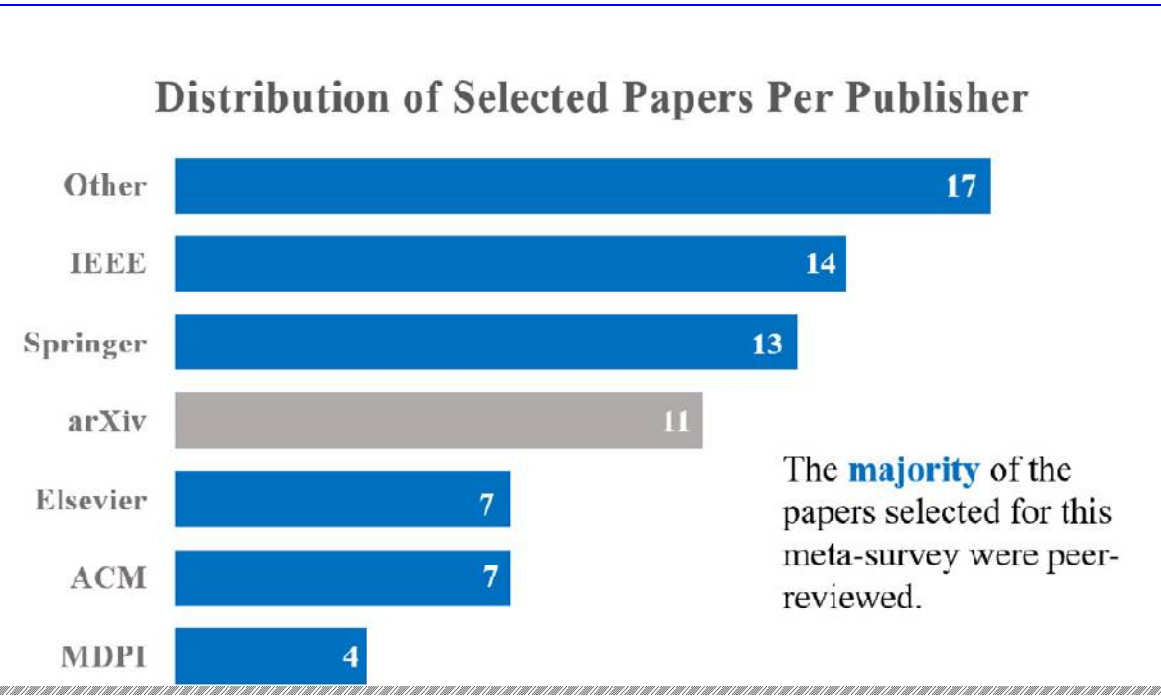
Last 10-year trends in academic publications in xAI



- ✓ Last 10-year trends in academic production (blue line)
- ✓ Articles indexed in: Elsevier Scopus database, and generic publishing (red line)

xAI.		2023-123


Distribution of Selected Papers Per Publication Type
Book Chapters
(5 papers)
7%
Book (1 book)
1%
arXiv Papers
(11 papers)
15%
Journal Articles
(41 papers)
56%
Conference
Papers
(15 papers)
21%
The primary outlet for the selected papers is journal articles followed by conference papers and arXiv papers.

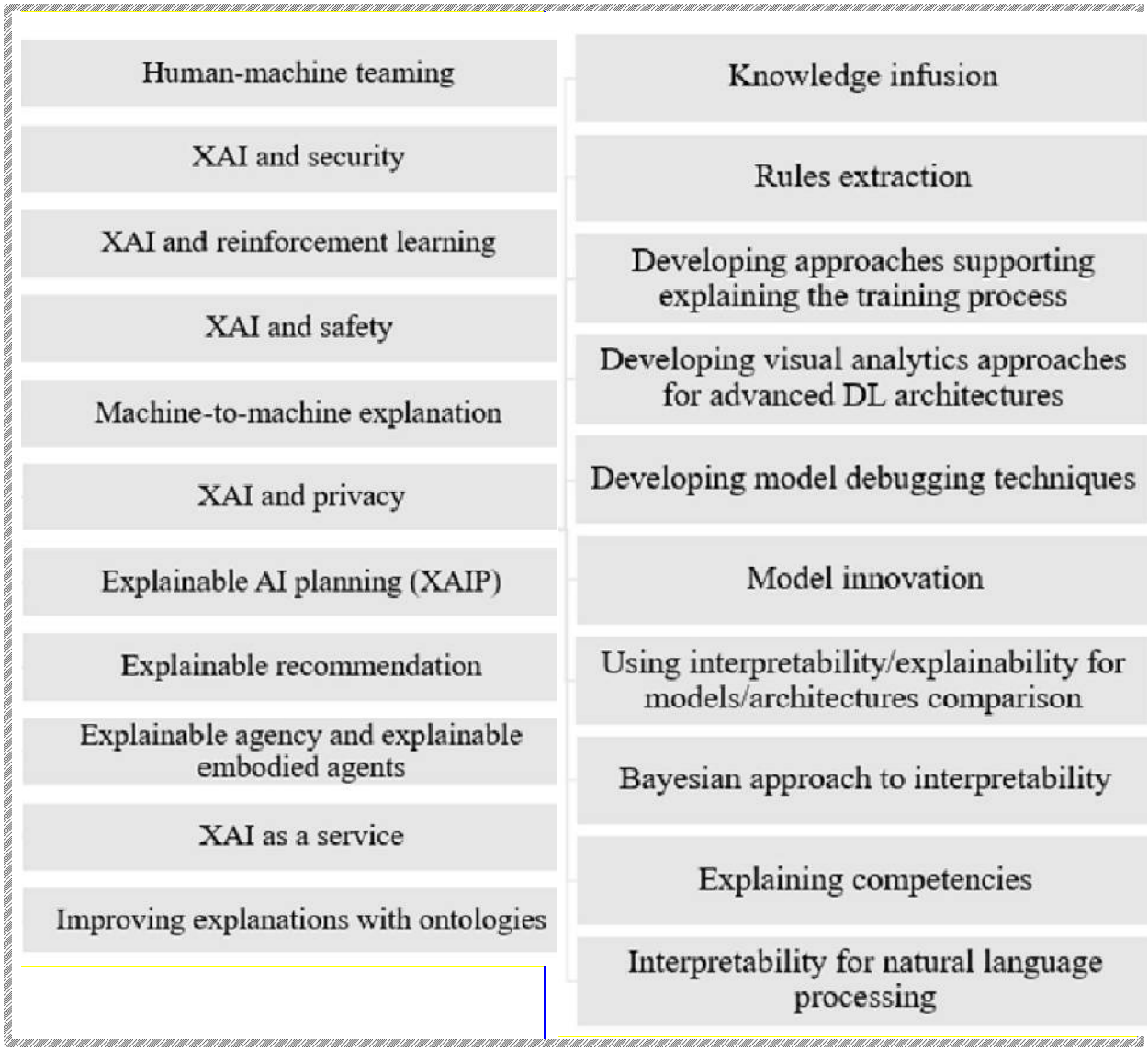

Distribution of Selected Papers Per Publisher
Other
17
IEEE
14
Springer
13
arXiv
11
Elsevier
7
ACM
7
MDPI
4
The majority of the papers selected for this meta-survey were peer-reviewed.

Recent research published in 2020 and 2021 has focused on most of these challenges and research directions

Challenges and research directions	2017	2018	2019	2020	2021	Total
Towards more formalism	1	4	4	11	11	31
Explanations and the nature of user experience and expertise	1	1	2	6	2	12
XAI for trustworthiness AI	1	1	0	4	5	11
Multidisciplinary research collaborations	0	1	1	3	5	10
Interpretability vs. performance trade-off	0	1	1	3	2	7
XAI for non-image, non-text, and heterogeneous data	0	0	0	3	3	6
Explainability methods composition	0	3	0	1	1	5
Causal explanations	0	1	0	3	1	5
Challenges in the existing XAI models/methods	0	1	1	1	2	5
Contrastive and counterfactual explanations	0	0	1	0	2	3
Communicating uncertainties	0	0	0	2	0	2
Time constraints	1	0	0	1	0	2
Natural language generation	0	0	0	1	0	1
Analyzing assumption-free black-box models, not assumption-based data models	0	0	1	0	0	1
Reproducibility	0	0	0	1	0	1
The economics of explanations	0	1	0	0	0	1

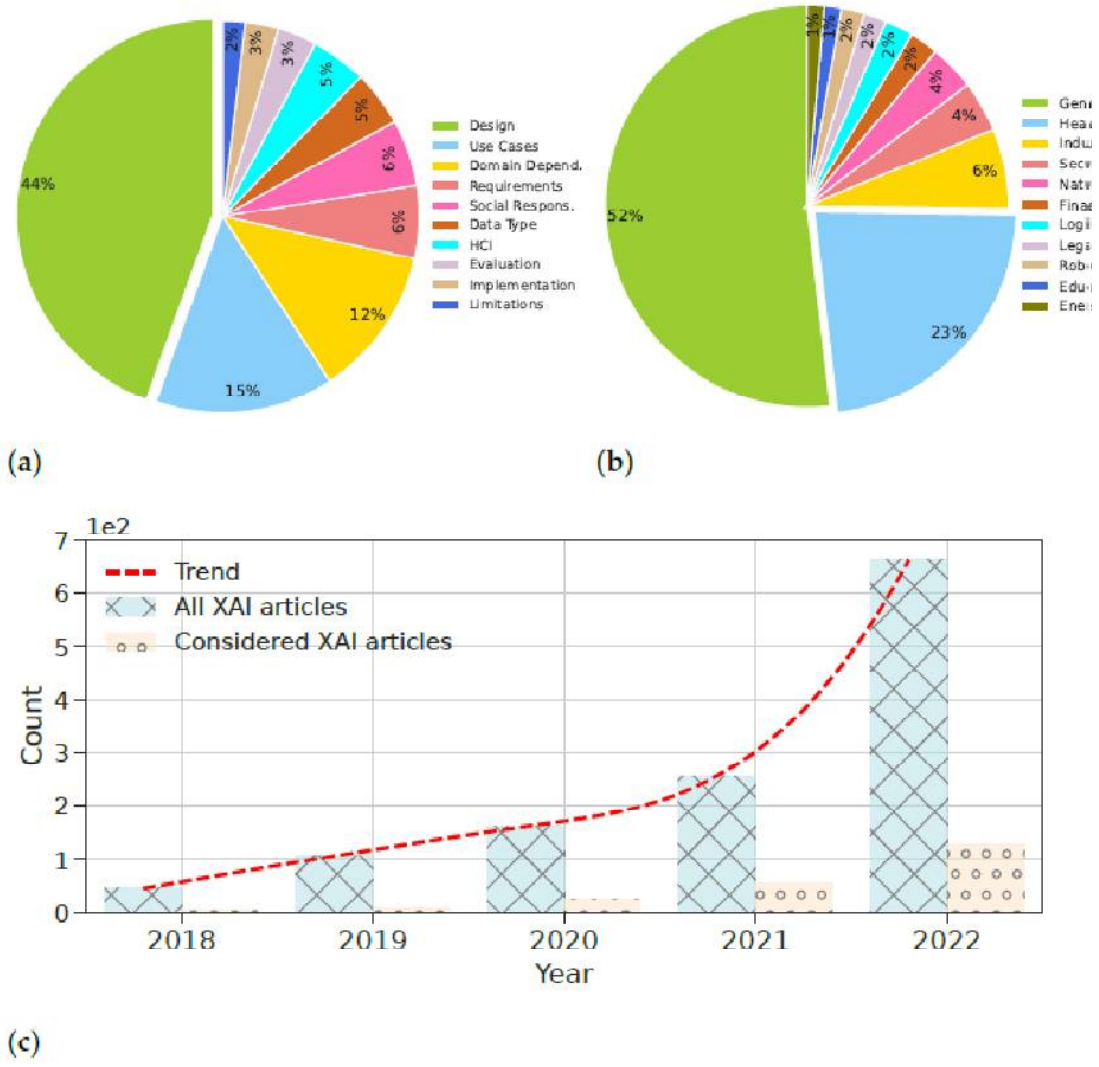
Recent research published in 2020 and 2021 has focused on most of these challenges and research directions

Phases	Challenges and research directions	2017	2018	2019	2020	2021	Total
Design							
	Communicating data quality	0	0	0	0	2	2
	Sparsity of analysis	0	0	0	1	0	1
Development							
	Knowledge infusion	0	2	0	4	0	6
	Rules extraction	0	1	0	1	2	4
	Developing approaches supporting explaining the training process	0	1	1	2	0	4
	Developing visual analytics approaches for advanced DL architectures	0	1	0	1	0	2
	Developing model debugging techniques	0	1	0	1	0	2
	Model innovation	0	0	0	0	2	2
	Using interpretability/explainability for models/architectures comparison	0	0	0	1	0	1
	Bayesian approach to interpretability	1	0	0	0	0	1
	Explaining competencies	0	0	1	0	0	1
	Interpretability for natural language processing	0	0	0	0	1	1
Deployment							
	Human-machine teaming	0	4	2	3	3	12
	XAI and security	0	1	0	3	3	7
	XAI and reinforcement learning	0	1	1	1	3	6
	XAI and safety	0	0	0	3	1	4
	Machine-to-machine explanation	0	2	1	0	0	3
	XAI and privacy	0	1	0	1	0	2
	Explainable AI planning (XAIP)	1	1	0	0	0	2
	Explainable recommendation	0	0	0	2	0	2
	Explainable agency and explainable embodied agents	0	0	1	0	1	2
	XAI as a service	0	0	1	0	1	2
	Improving explanations with ontologies	0	0	0	0	1	1



xAI.		2023-145

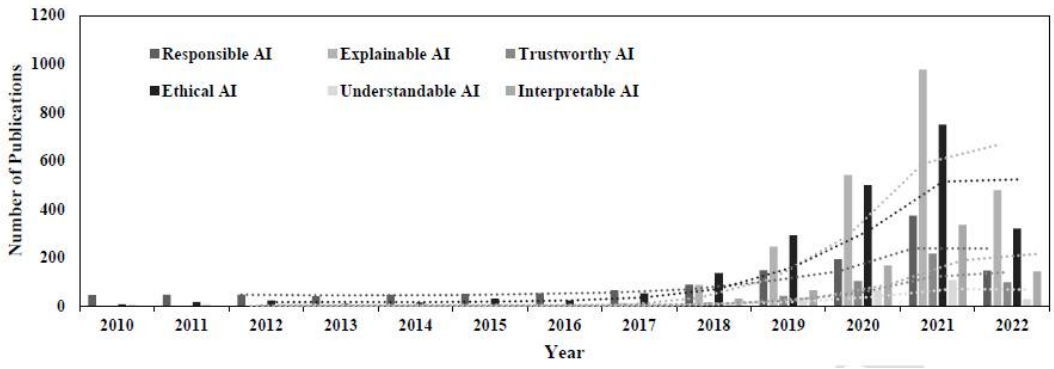
Research publications -- xAI



(a) Research focus. (b) Application domains (c) Time distribution.

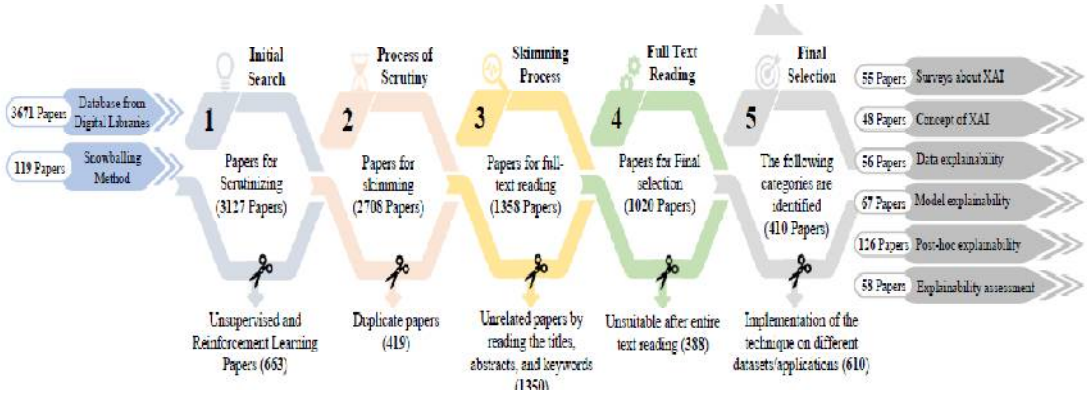
xAI.		2023-030

Total number of publications on XAI over time



- ✓ Dotted lines: trend over the previous three years using a moving average
- ✓ Statistics were retrieved from the Scopus database in June 2022

Search strategy for categorizing selected XAI research papers on supervised learning



Keyword Search

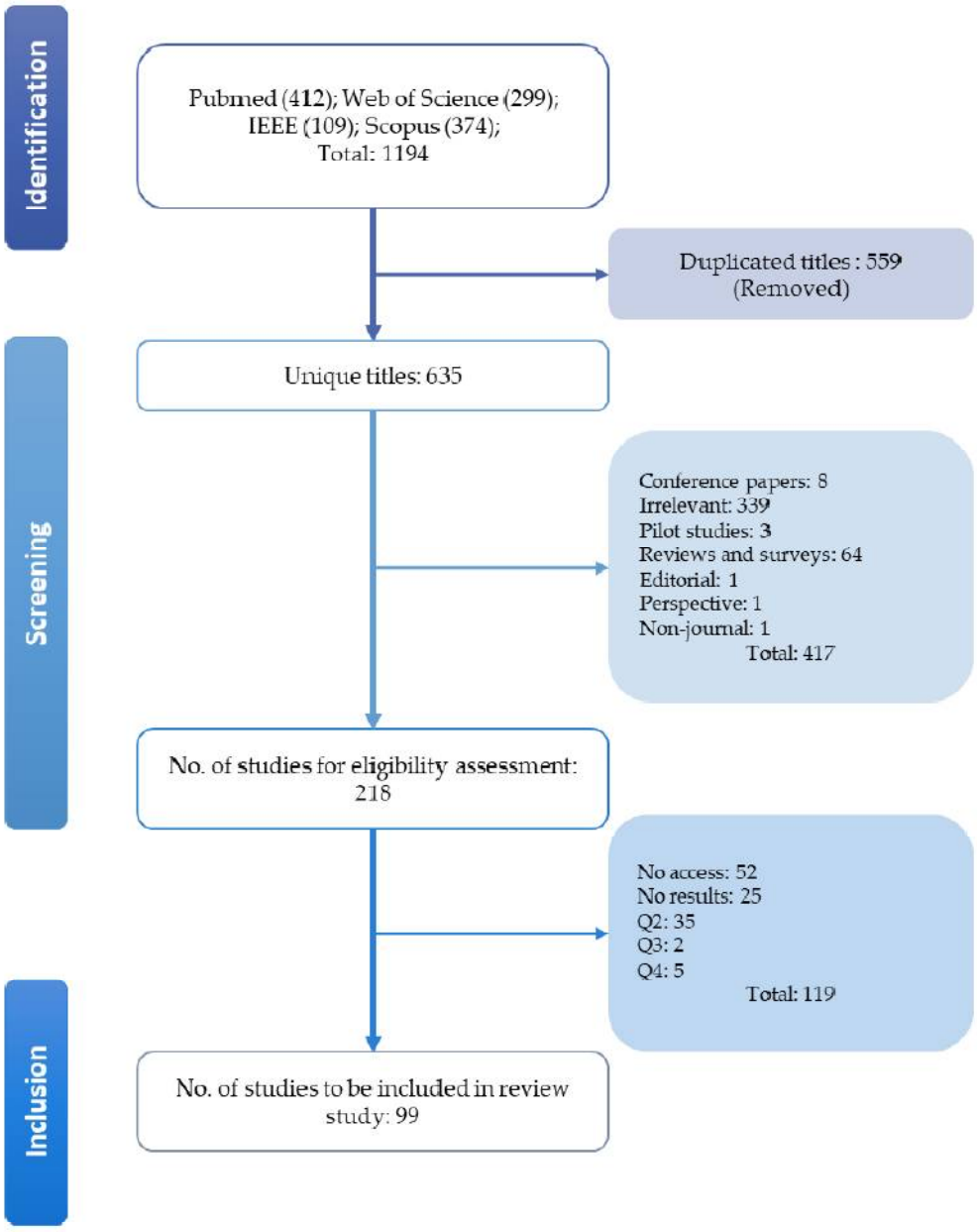
xAI.		2023-150

Boolean search strings used for journal databases

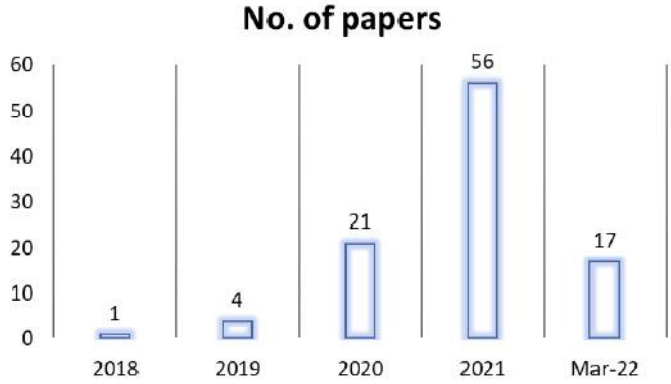
Database	Boolean search strings
Scopus	(TITLE-ABS (explainability) OR TITLE-ABS (explainable)) AND (TITLE-ABS (machine AND learning) OR TITLE-ABS (deep AND learning)) AND (TITLE-ABS (medical) OR TITLE-ABS (clinical) OR TITLE-ABS (health) OR TITLE-ABS (healthcare) OR TITLE-ABS (biomedical)) AND (LIMIT-TO (DOCTYPE , "ar"))
PubMed	(explainability[Title/Abstract] OR explainable[Title/Abstract]) AND (machine learning[Title/Abstract] OR deep learning [Title/Abstract]) AND (medical[Title/Abstract] OR biomedical[Title/Abstract] OR health[Title/Abstract] OR healthcare[Title/Abstract] OR clinical[Title/Abstract])
Web of Science	(AB=(explainable) OR AB=(explainability)) AND (AB=(machine learning) OR AB=(deep learning)) AND (AB=(medical) OR AB=(biomedical) OR AB=(health) OR AB=(healthcare) OR AB=(clinical))
IEEE	(("Abstract":"explainable" OR "Abstract":"explainability") AND ("Abstract":"machine learning" OR "Abstract":"deep learning"))

xAI.		2023-150

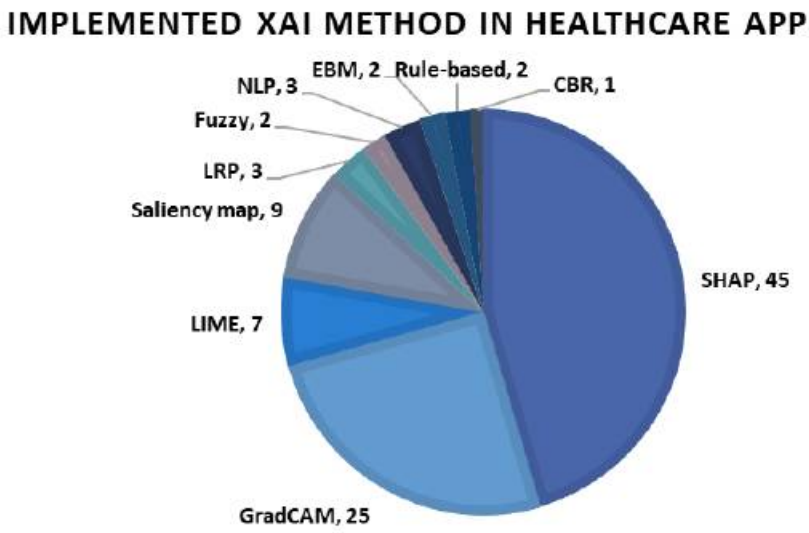
PRISMA systematic filtration of journal articles



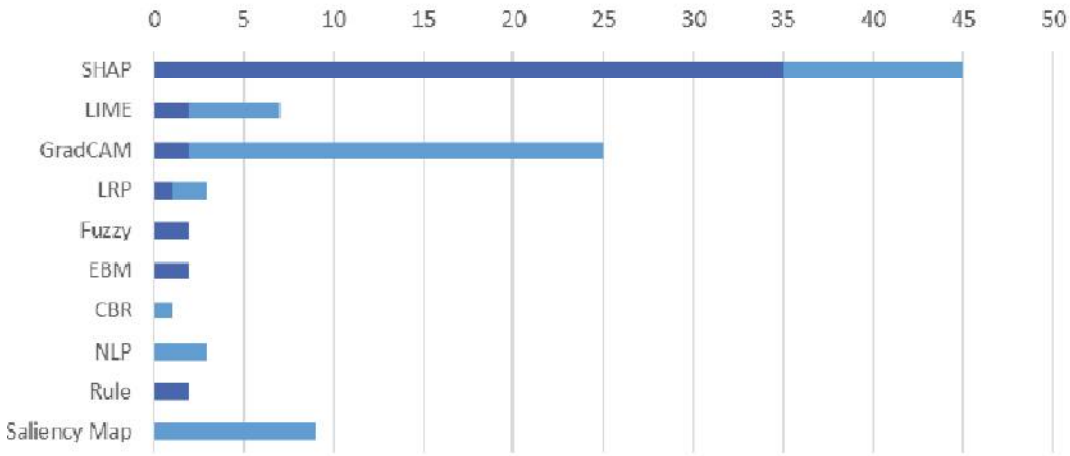
Bar chart representation of XAI studies from 2018 to March 2022



Pie chart of XAI techniques employed



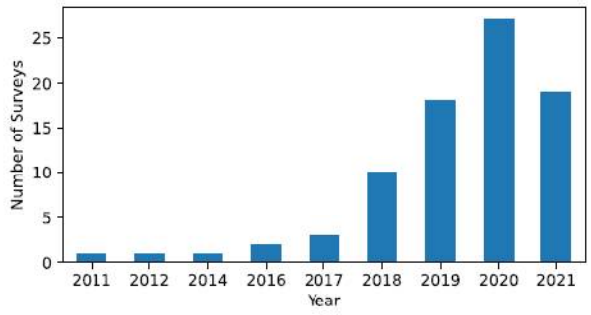
Bar chart representation of the number of ML and DL models used in conjunction with each XAI technique

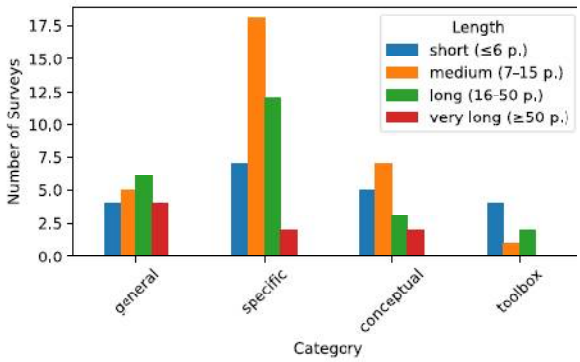


	Saliency Map	Rule	NLP	CBR	EBM	Fuzzy	LRP	GradCAM	LIME	SHAP
ML	0	2	0	0	2	2	1	2	2	35
DL	9	0	3	1	0	0	2	23	5	10

Distribution of the reviewed surveys over years 2010 to 2021

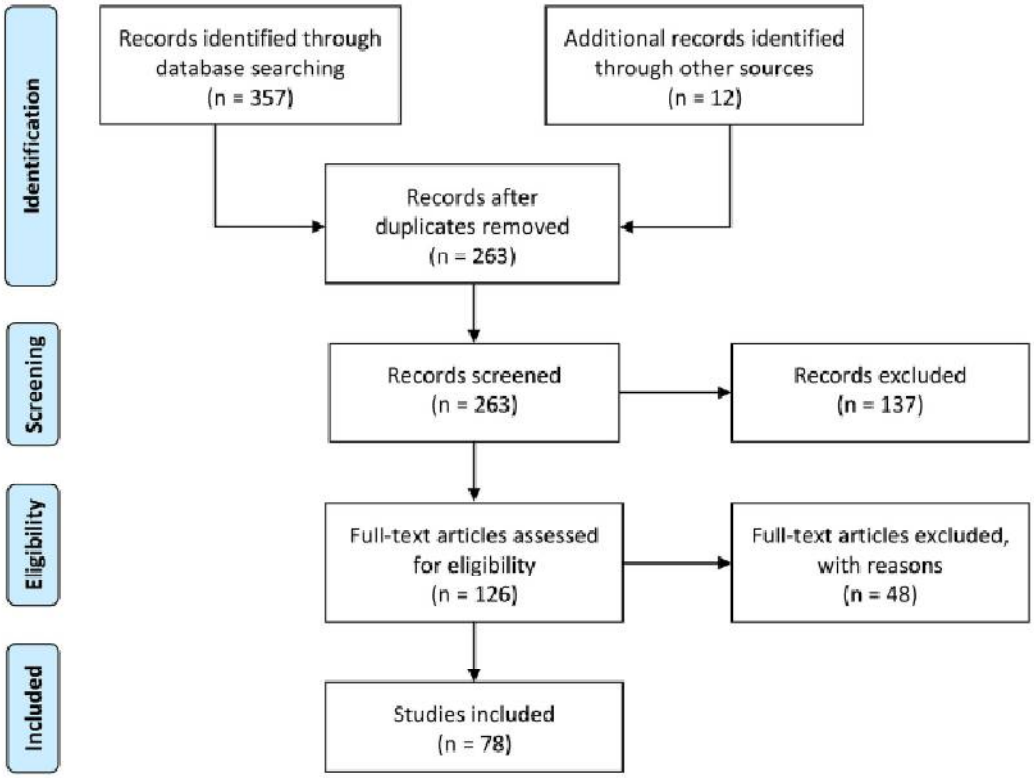
Distribution of the survey Lengths by general focus category





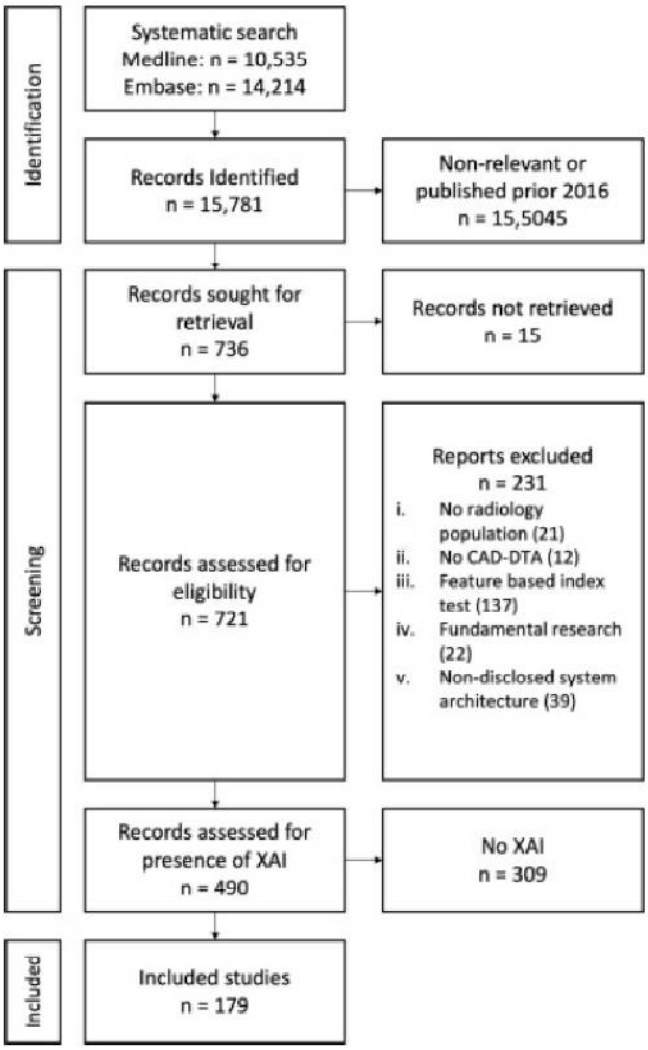
xAI. 2022-103

PRISMA flow diagram
for selection processes in systematic review

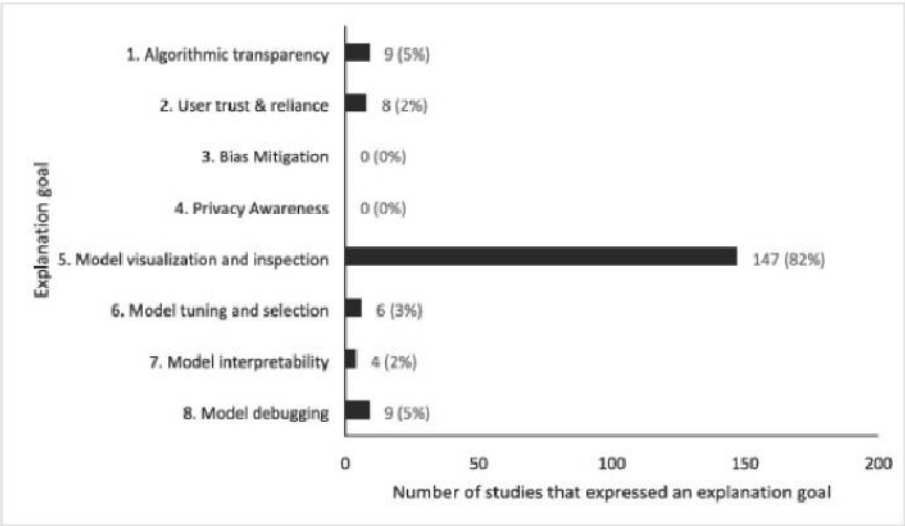


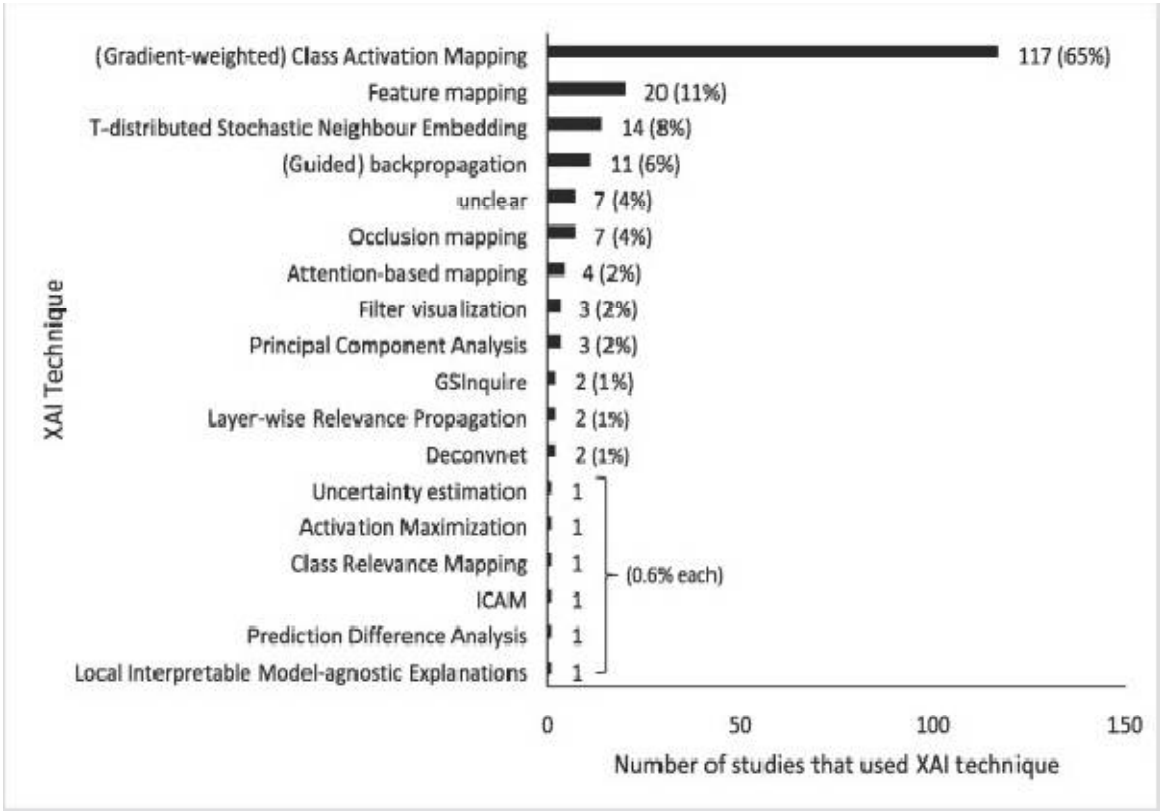
xAI. 2022-066

PRISMA flow diagram

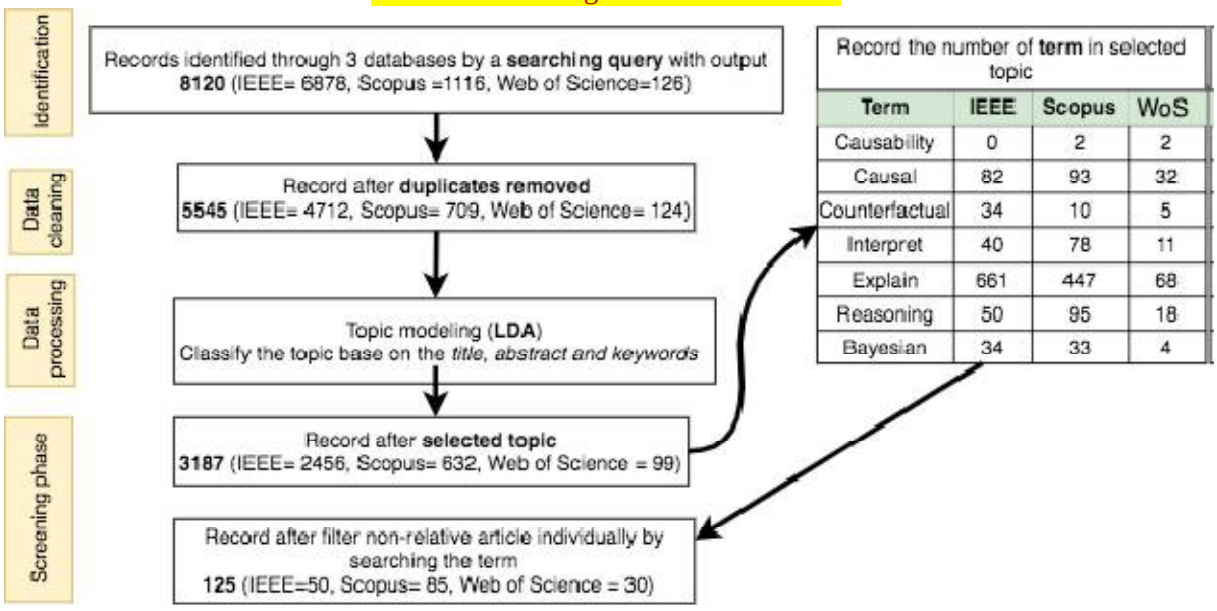


Explanation goals



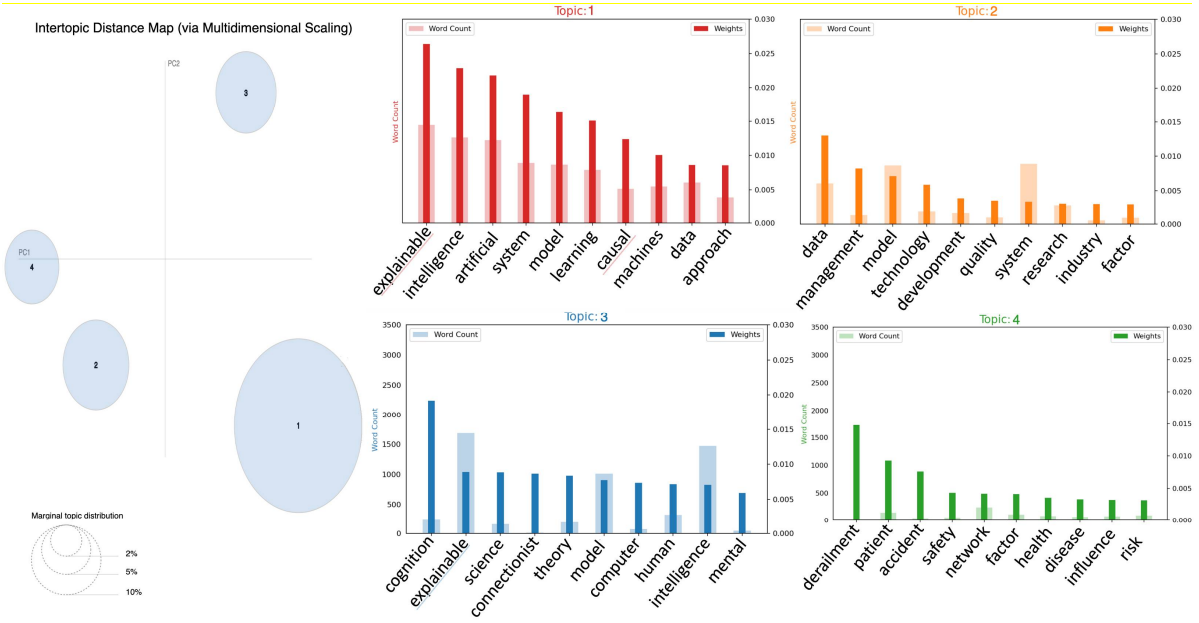


PRISMA flow diagram & search results

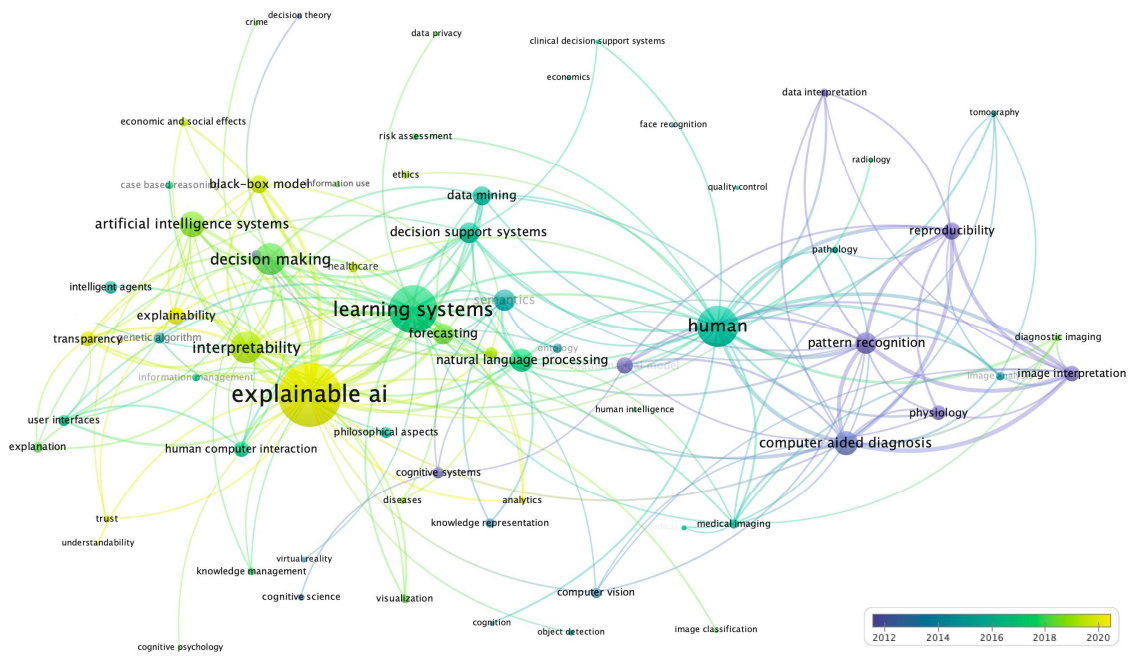


Record the number of **term** in selected topic

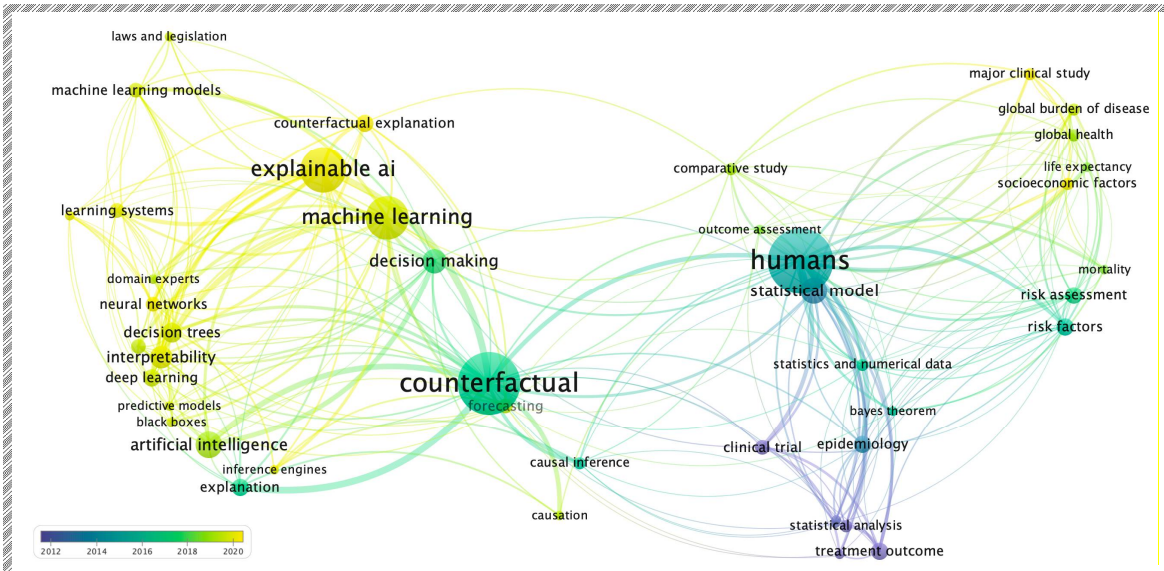
Term	IEEE	Scopus	WoS
Causability	0	2	2
Causal	82	93	32
Counterfactual	34	10	5
Interpret	40	78	11
Explain	661	447	68
Reasoning	50	95	18
Bayesian	34	33	4



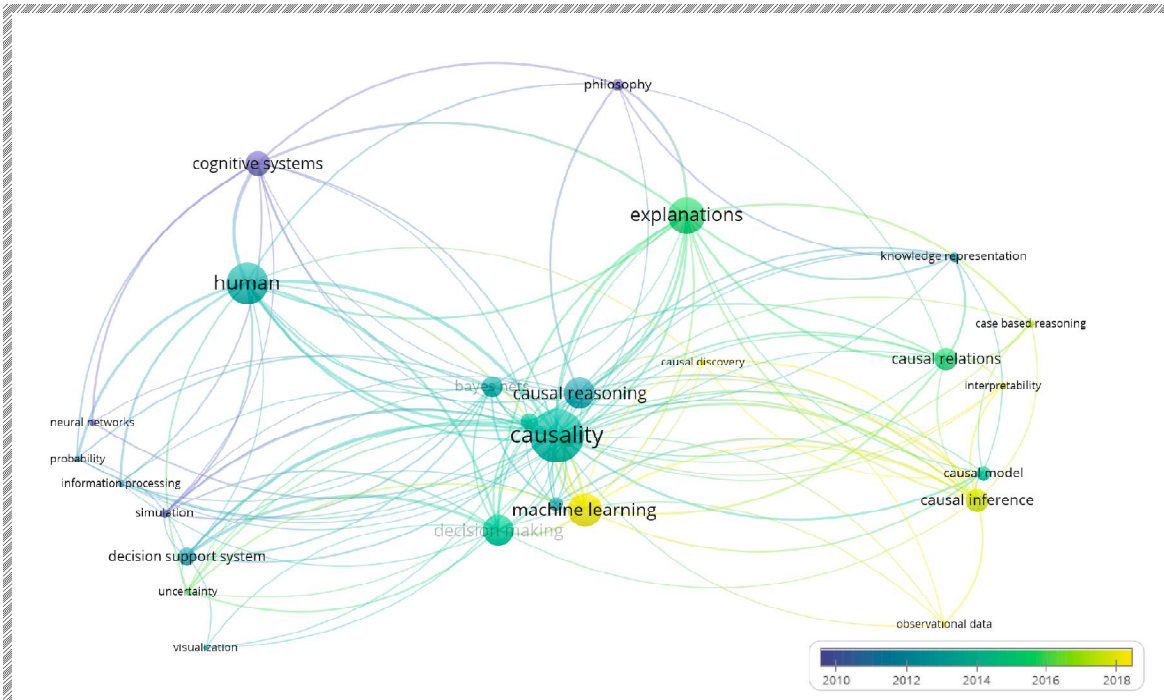
- ✓ Key words: LDA: Scopus database
- ✓ 709 titles, abstracts, and authors keywords
- ✓ Figure also shows the top 10 most relevant words for each Topic



Network visualization of co-occurrence between keywords in articles about XAI



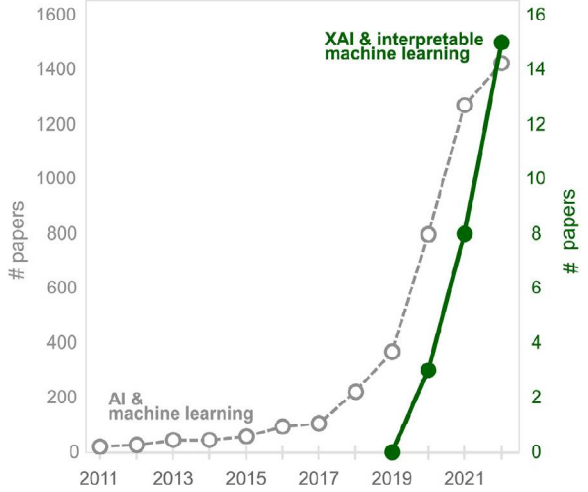
Network visualization of co-occurrence between keywords in articles about counterfactuals in XAI



Network visualization of co-occurrence between keywords in articles about causality in XAI

xAI.		2022-153

"AI and machine learning"
"XAI and interpretable machine learning"
in agricultural science



✓ Web of Science Core Collection

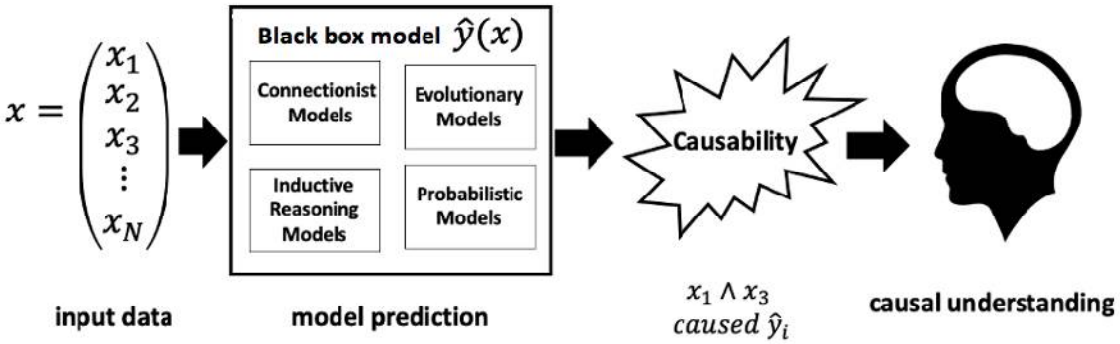
Black--To--white box through grey box AI methods

xAI. | 2022-13

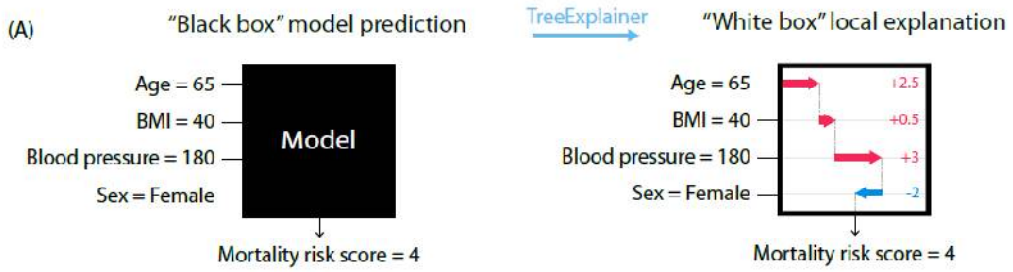
Inp-BBM-Out

xAI. | 2022-143

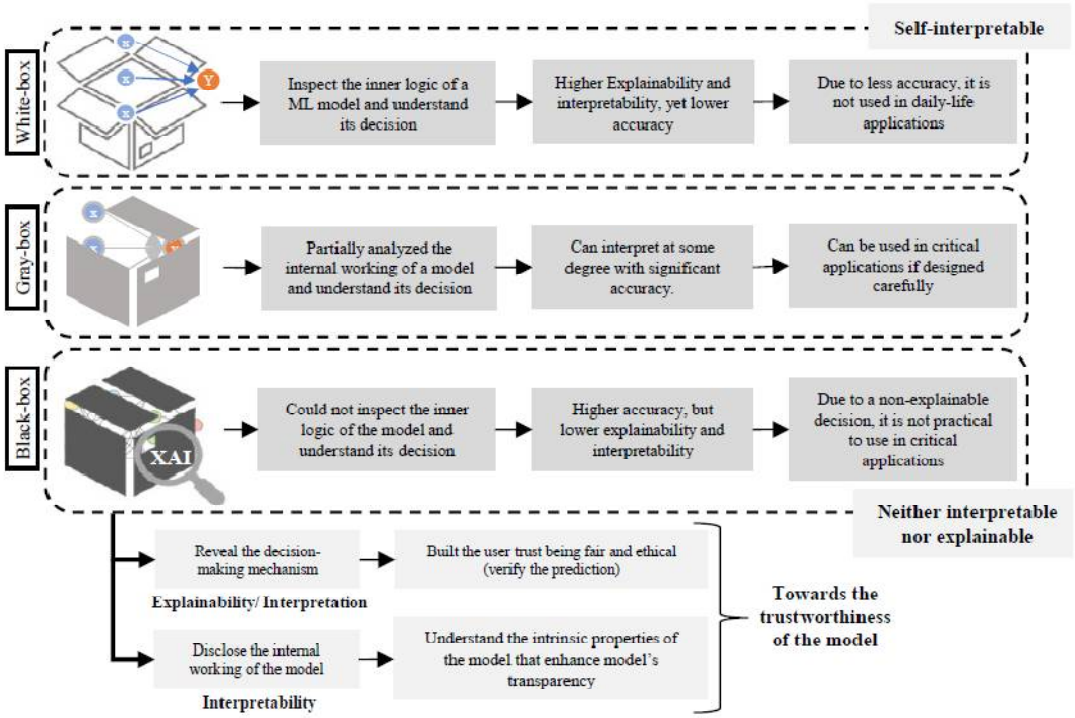
Black box—Causality explanation



Black—White box Methods



White-box, gray-box, black-box models



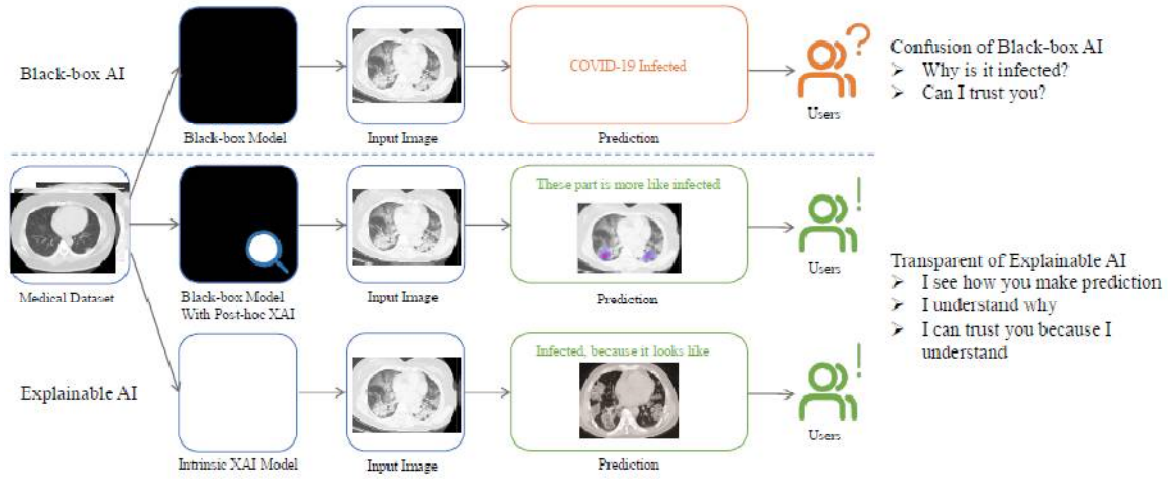
- Black-box models are more accurate but less interpretable
- White-box models are interpretable by design. → making their outputs easier to understand

but less accurate.

- Gray-box models yield a good interpretability/accuracy tradeoff
- Future: More complex XAI techniques are required for creating trustworthy models.

COVID detection

Comparison between black-box and explainable artificial intelligence



- Top branch shows the process of a black-box model. It provides only results such as classes (e.g., COVID or non-COVID).
- Middle and bottom branches: Two XAI methods;
- Middle branch: Example of saliency map
- Bottom branch: Prototype method

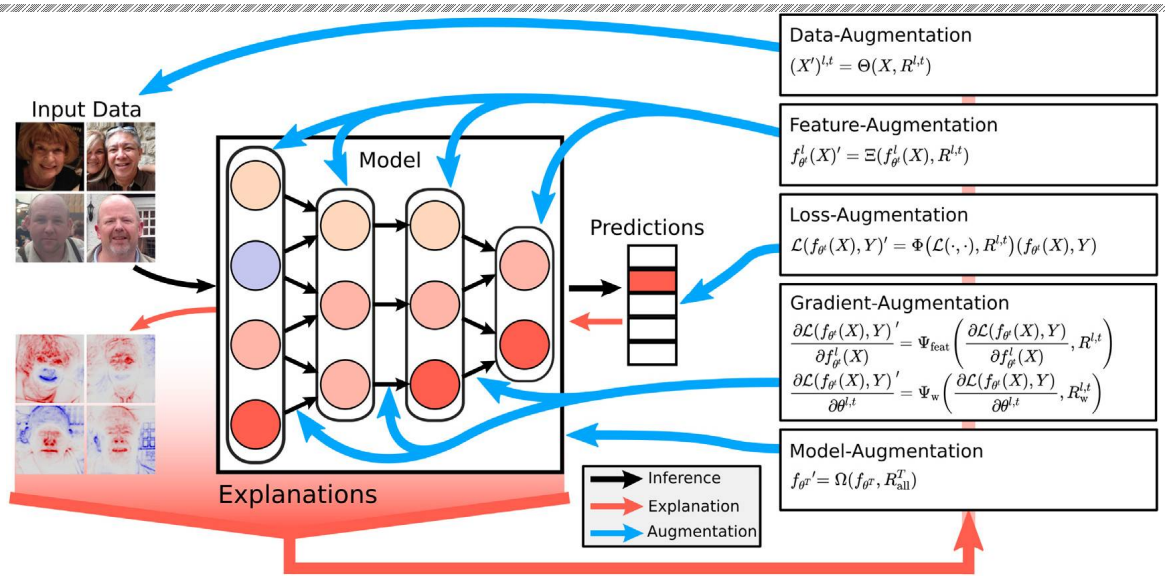
xAI. 2023-

Black-box to white box xAI

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

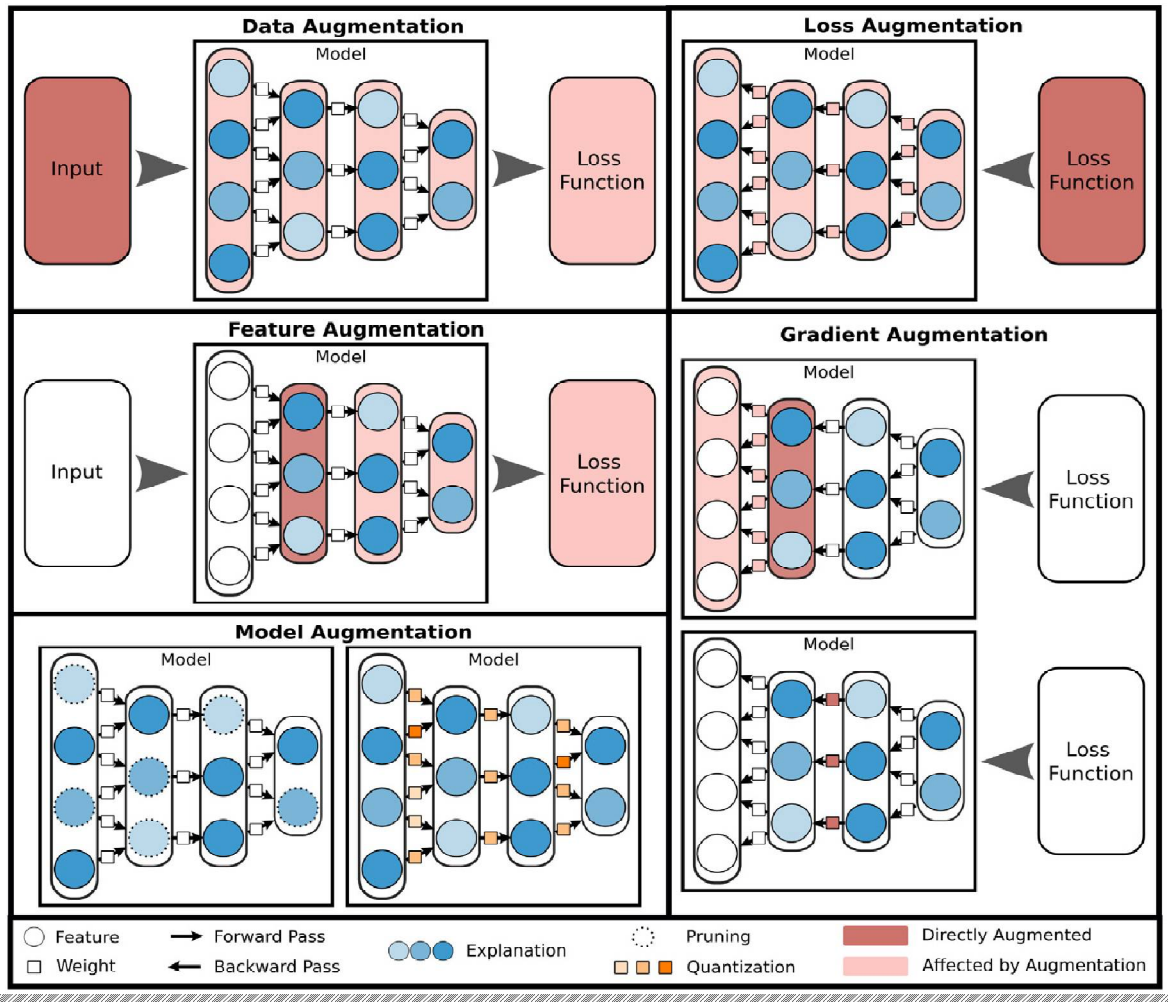
xAI. 2023-079

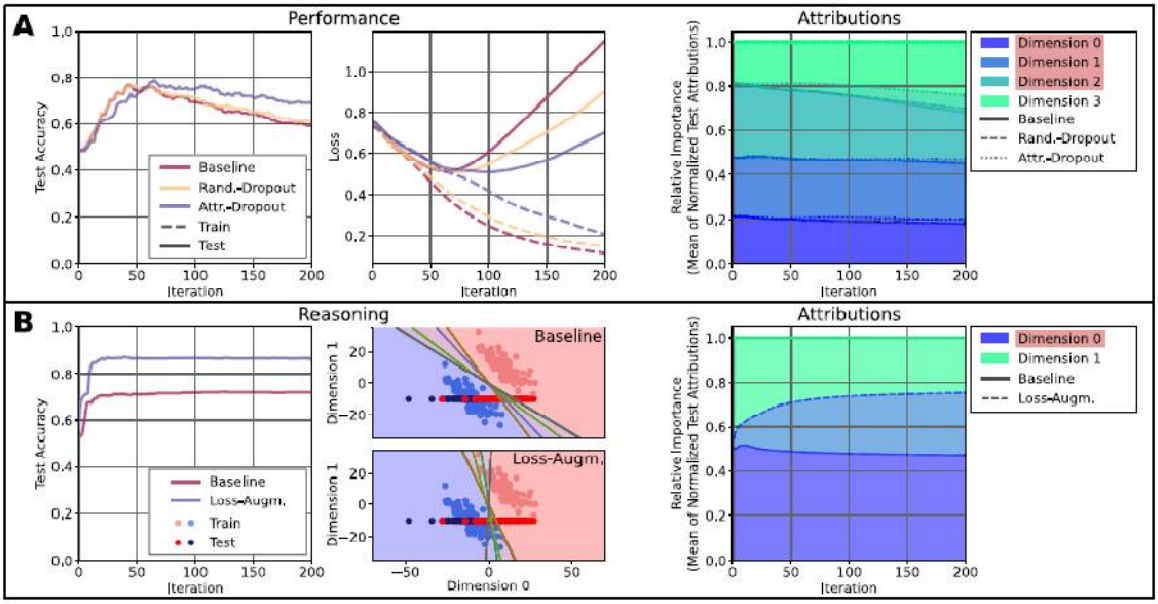
Model improvement with XAI



✓ Explanations offer information about the model decision-making and behavior → used to improve models by augmenting different components of the training process

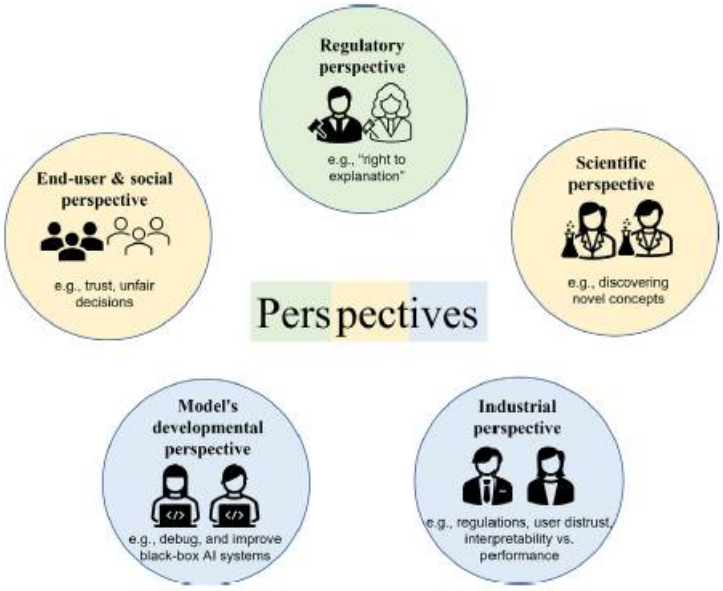
Types of XAI-based augmentation





xAI.		2023-123

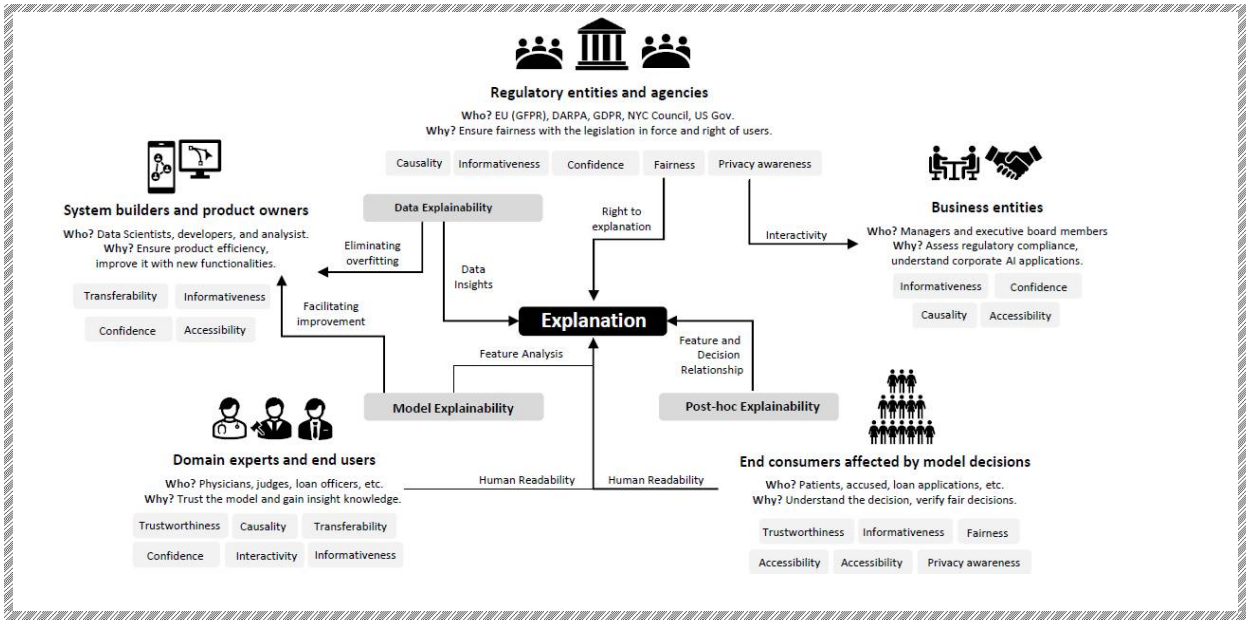
xAI- perspectives



xAI.		2023-030

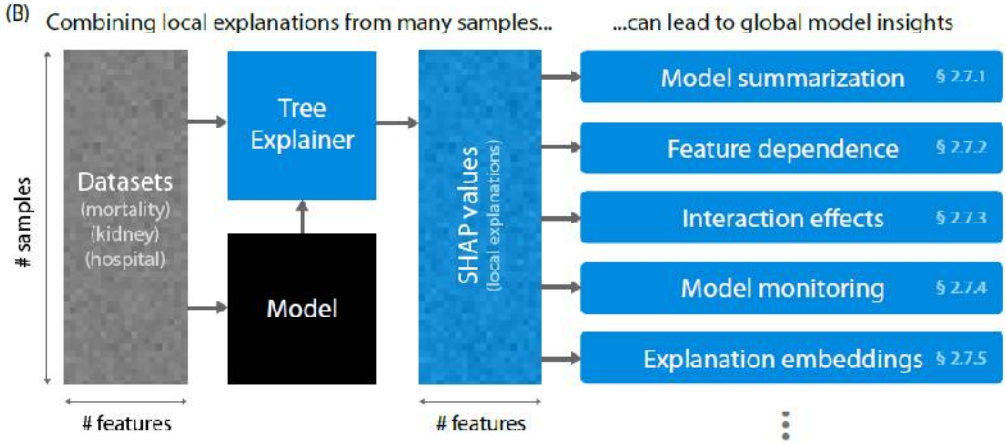
xAI stakeholders

(End-users, domain experts, developers, government bodies)



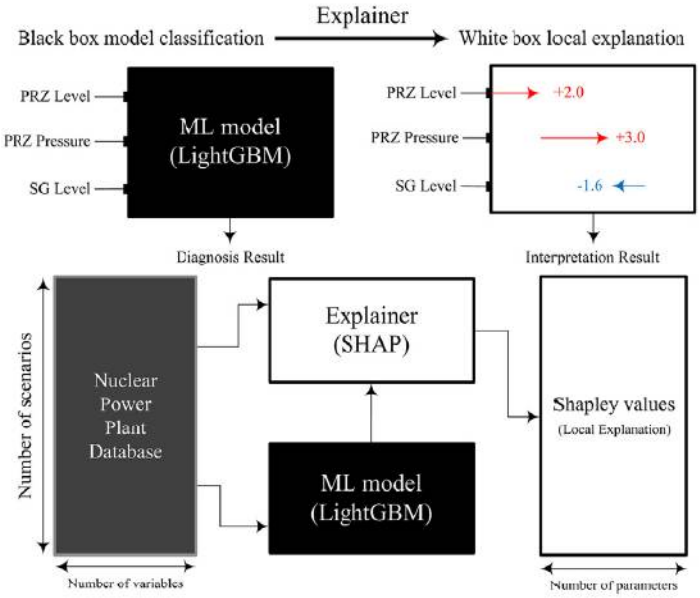
xAI. | 2023-158

Local/Global explanations

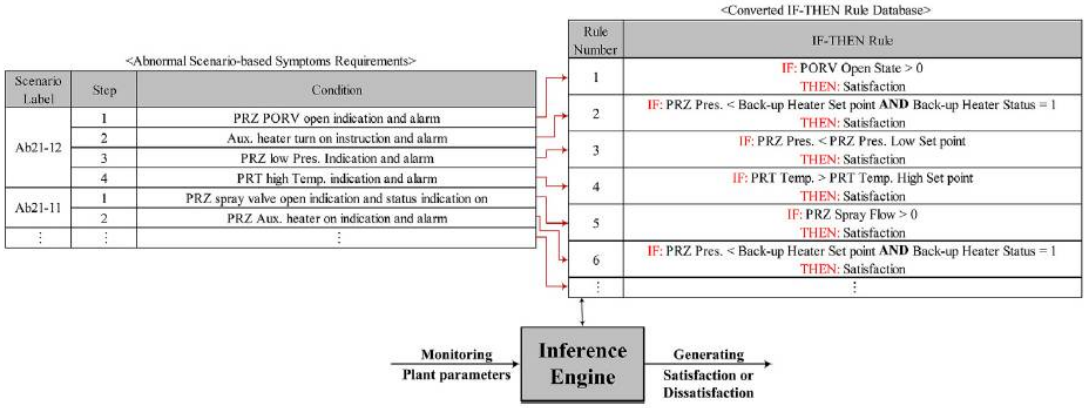


xAI. | 2022-159

xAI



Structure of the rule-based system



<Abnormal Scenario-based Symptoms Requirements>

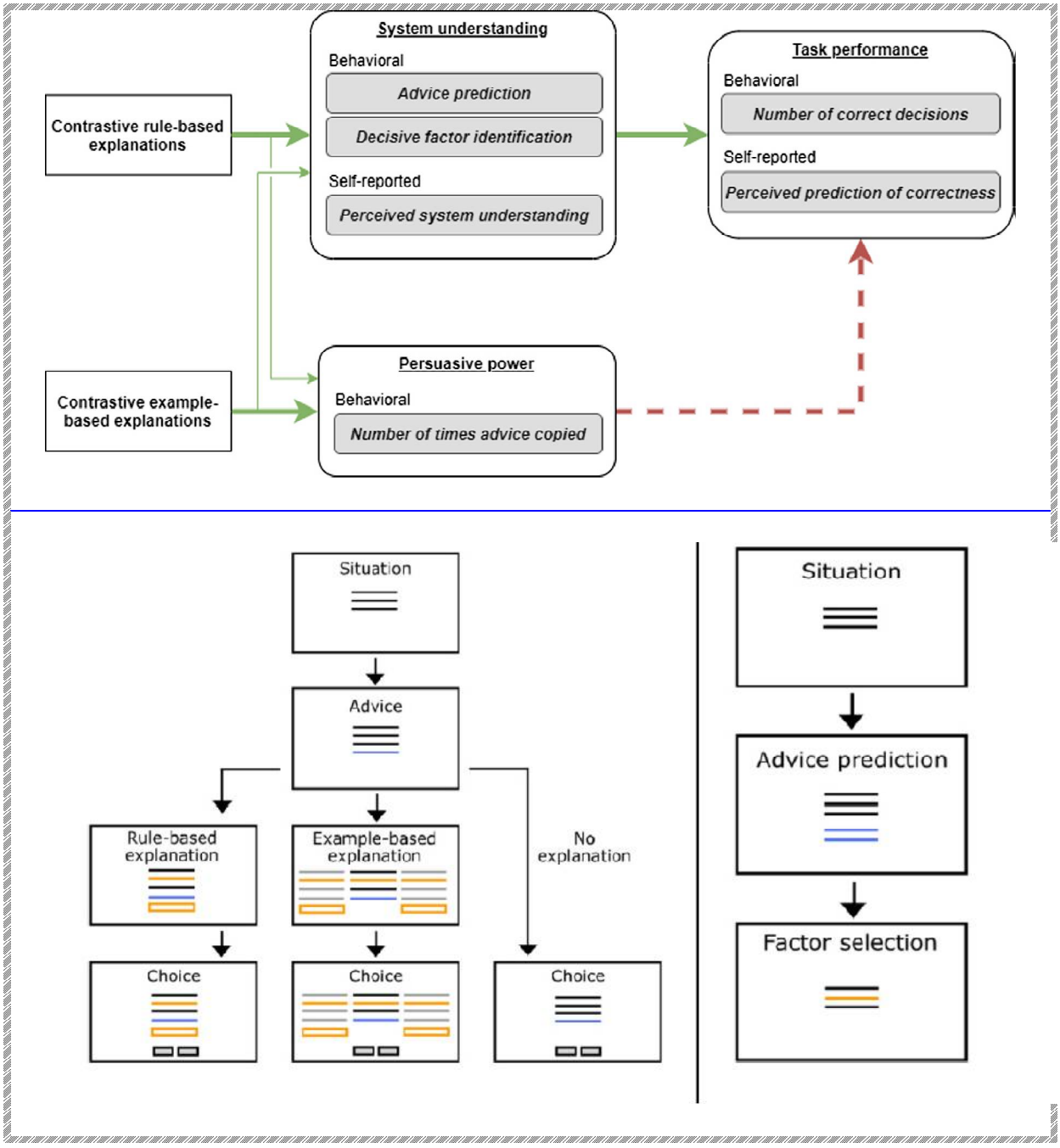
Scenario Label	Step	Condition
Ab21-12	1	PRZ PORV open indication and alarm
	2	Aux. heater turn on instruction and alarm
	3	PRZ low Pres. Indication and alarm
	4	PRT high Temp. indication and alarm
Ab21-11	1	PRZ spray valve open indication and status indication on
	2	PRZ Aux. heater on indication and alarm
⋮	⋮	⋮

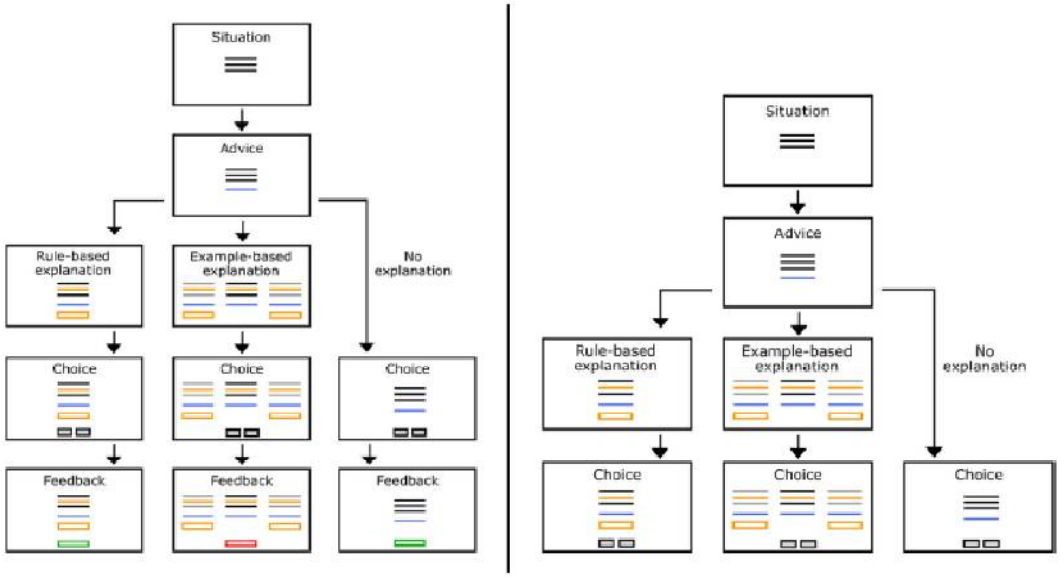
<Converted IF-THEN Rule Database>

Rule Number	IF-THEN Rule
1	IF: PORV Open State > 0 THEN: Satisfaction
2	IF: PRZ Pres. < Back-up Heater Set point **AND** Back-up Heater Status = 1 THEN: Satisfaction
3	IF: PRZ Pres. < PRZ Pres. Low Set point THEN: Satisfaction
4	IF: PRT Temp. > PRT Temp. High Set point THEN: Satisfaction
5	IF: PRZ Spray Flow > 0 THEN: Satisfaction
6	IF: PRZ Pres. < Back-up Heater Set point **AND** Back-up Heater Status = 1 THEN: Satisfaction
⋮	⋮

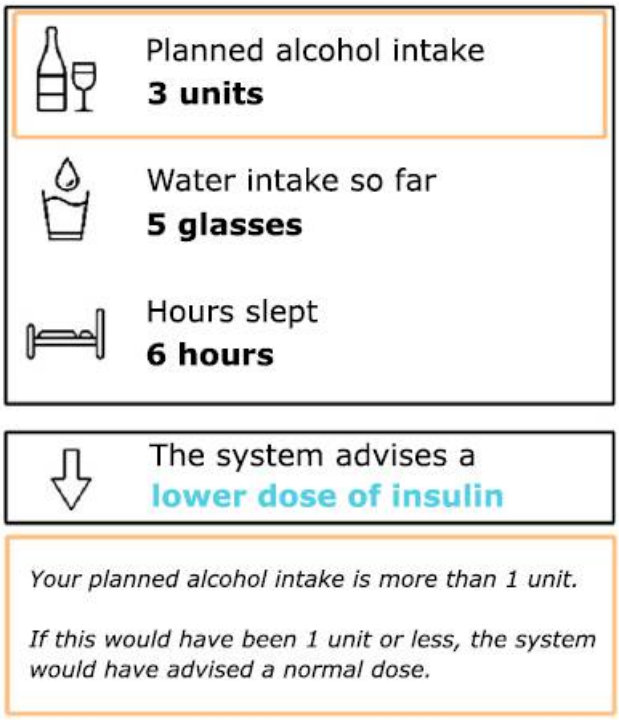
Explanation modes

xAI.		2023-149

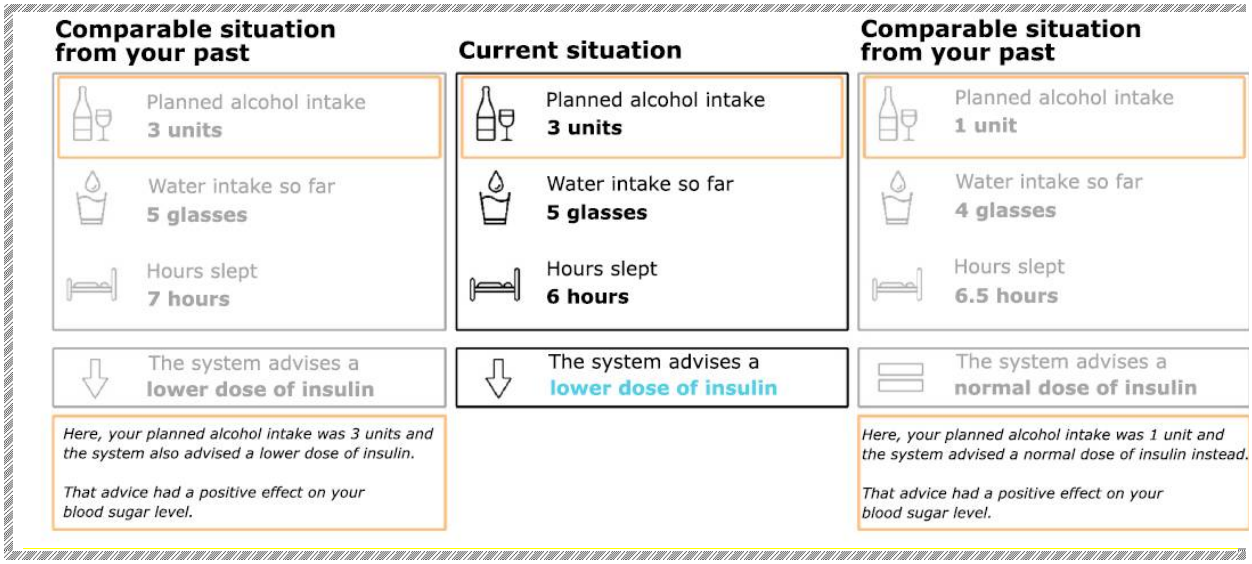

System understanding
Behavioral
Advice prediction
Decisive factor identification
Self-reported
Perceived system understanding
Task performance
Behavioral
Number of correct decisions
Self-reported
Perceived prediction of correctness
Contrastive rule-based explanations
Contrastive example-based explanations
Persuasive power
Behavioral
Number of times advice copied
Situation
Advice
Rule-based explanation
Example-based explanation
No explanation
Choice
Choice
Choice
Situation
Advice prediction
Factor selection

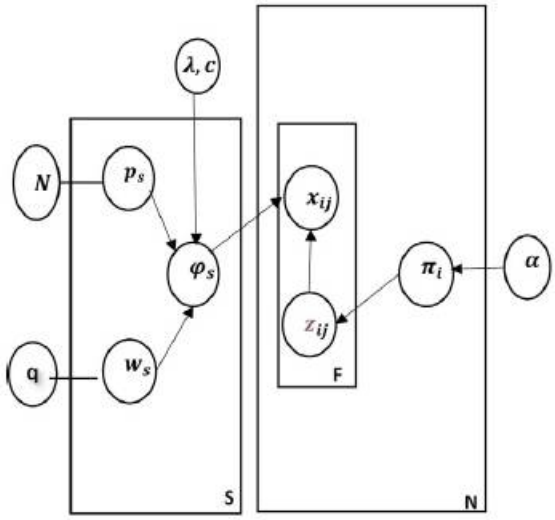


Rule-based explanation mode

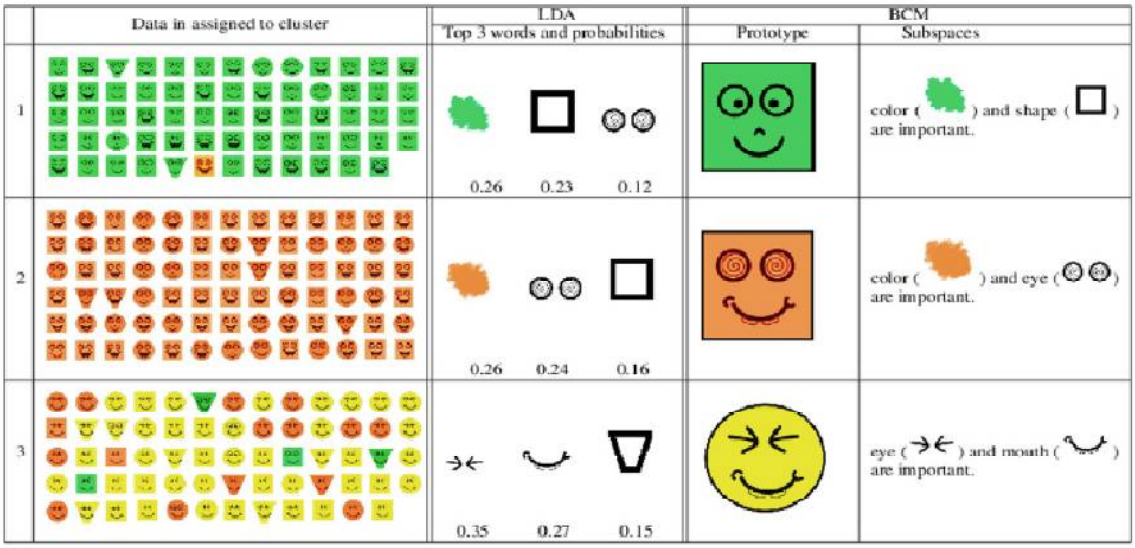


Example-based explanation styles





(a) Flow of Bayesian Case Model (BCM)

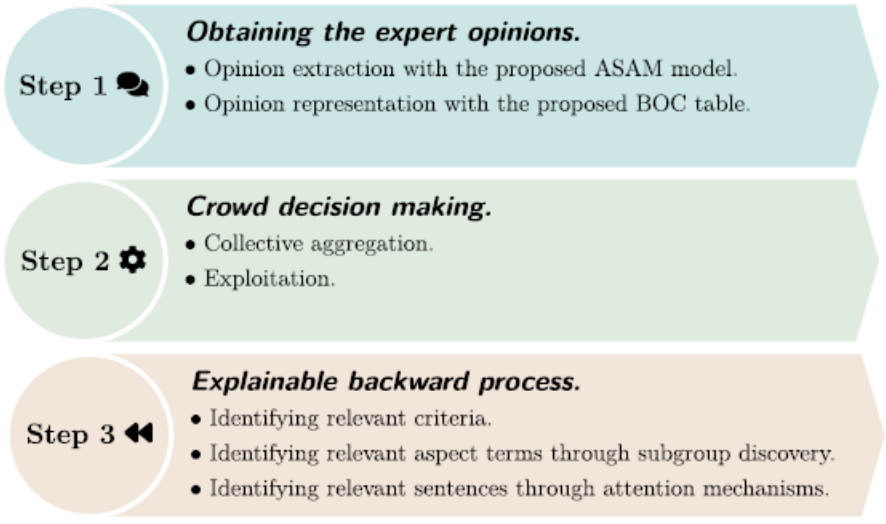


(b) BCM and LDA interpretation for dataset of smiley faces [77]

Workflows

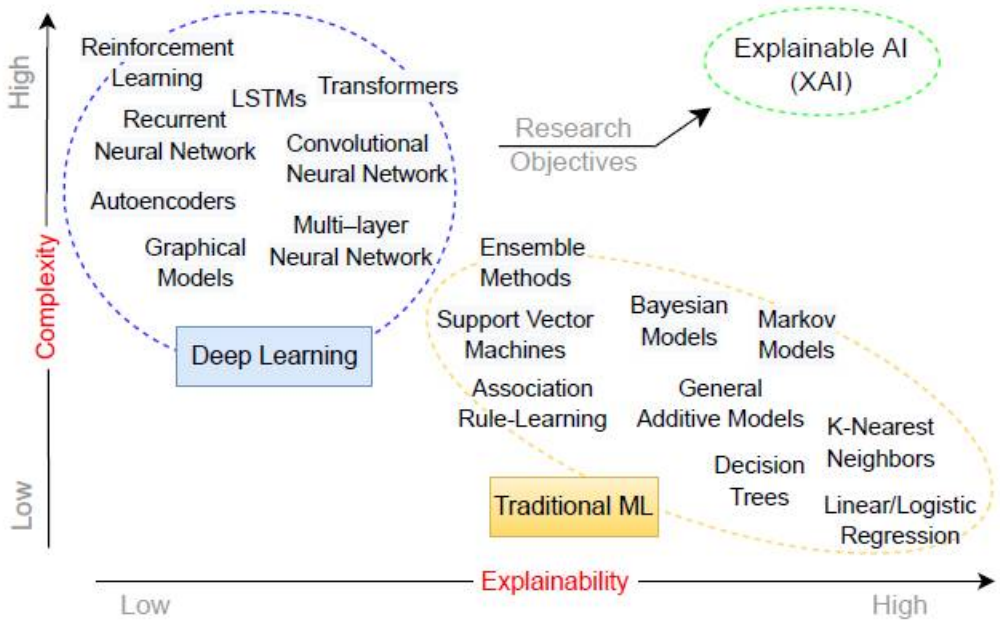
xAI.		2023-144

Workflow of ECDM-SDAM methodology



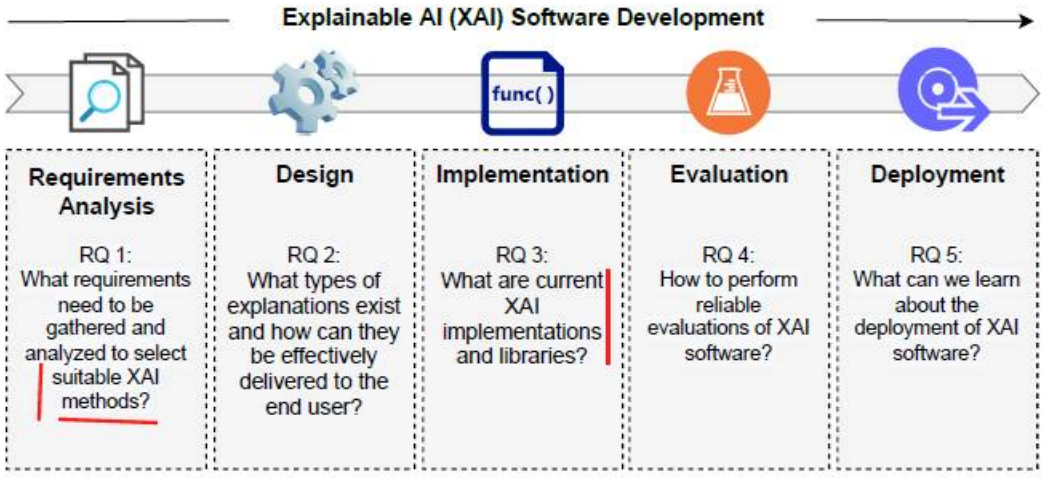
- Model: Crowd Decision Making (CDM)
- ECDM: Explainable CDM based on
- Methodology : Subgroup Discovery and Attention Mechanisms (SDAM)
- ASAM: Attention based Sentiment Analysis Method
- BOC : Bag of Opinions by Criteria

Classification of AI models

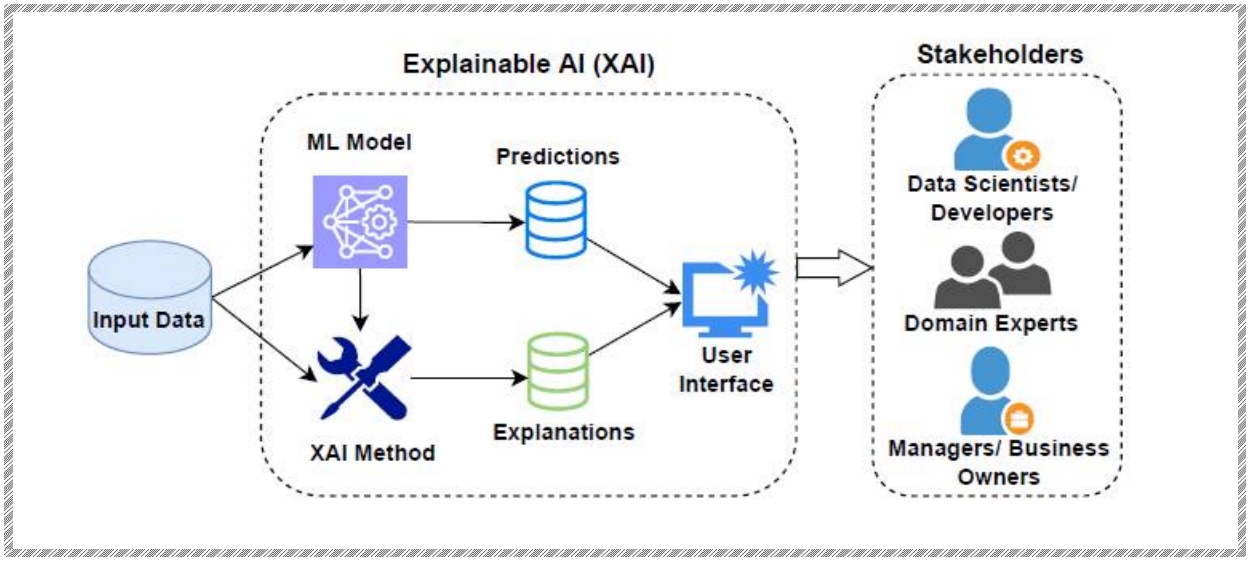


✓ Complexity, explainability, and potential applications

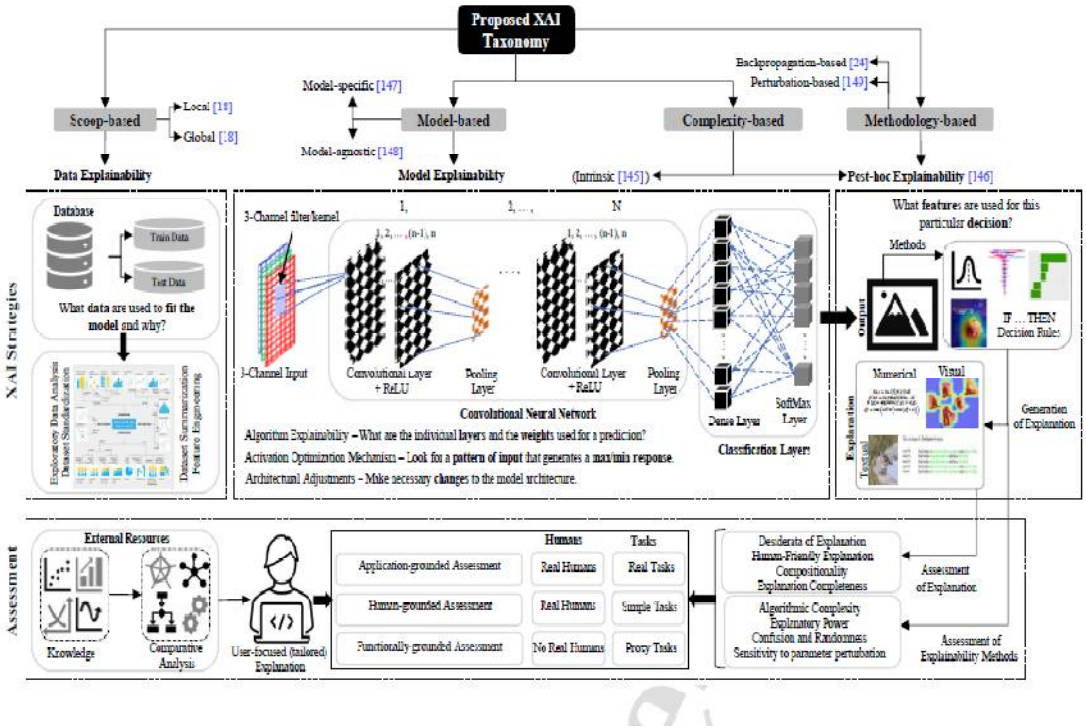
Addressed research questions in XAI software development process



Stakeholders for XAI



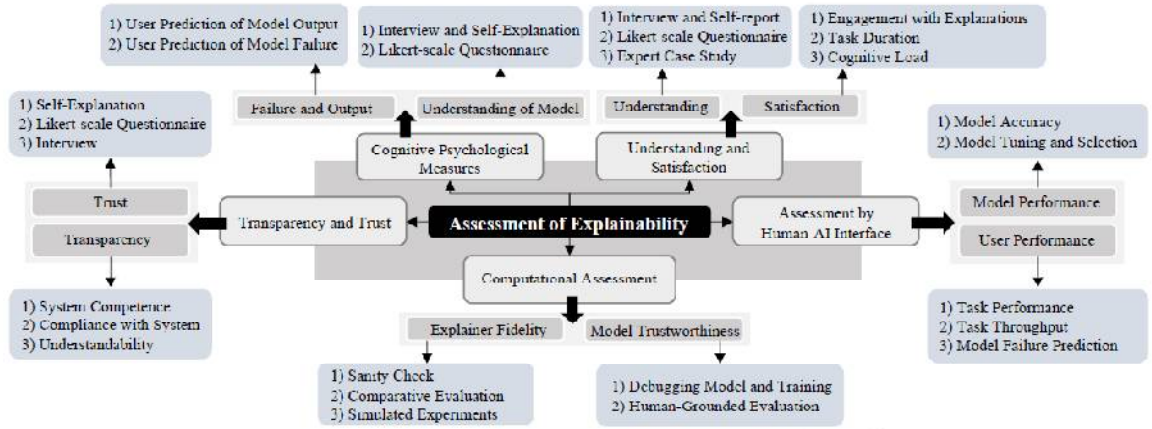
XAI methodology



By Data Explainability	By Model Explainability	By Post-hoc Explainability
D1: What sort of information do we have in the database? **D2:** What can be inferred from this data? **D3:** What are the most important portions of the data? **D4:** How is the information distributed? **D5:** Is it possible to increase the model's performance by lowering the number of dimensions? **D6:** Can a better explanation be offered by using data summarizing techniques?	**M1:** What makes a parameter, objective, or action important to the system? **M2:** When did the system examine a parameter, objective, or action, and when did the model reject it? **M3:** What are the consequences of making a different decision or adjusting a parameter? **M4:** How does the system carry out a certain action? **M5:** How do these model parameters, objectives, or actions relate to one another? **M6:** What factors does the system take into account (or disregard) when making a decision? **M7:** In order to achieve a goal/inference, which techniques does the system utilize or avoid?	**P1:** What is the reason behind the model's prediction? **P2:** What was the reason for occurrence X? What would happen if Y was the cause of occurrence X? **P3:** What variables have the most influence on the user's decision? **P4:** What if the information is altered? **P5:** To keep current results, what criteria must be met? **P6:** Is there anything that can be done to have a different outcome? **P7:** Why is it essential to make a certain conclusion or decision?

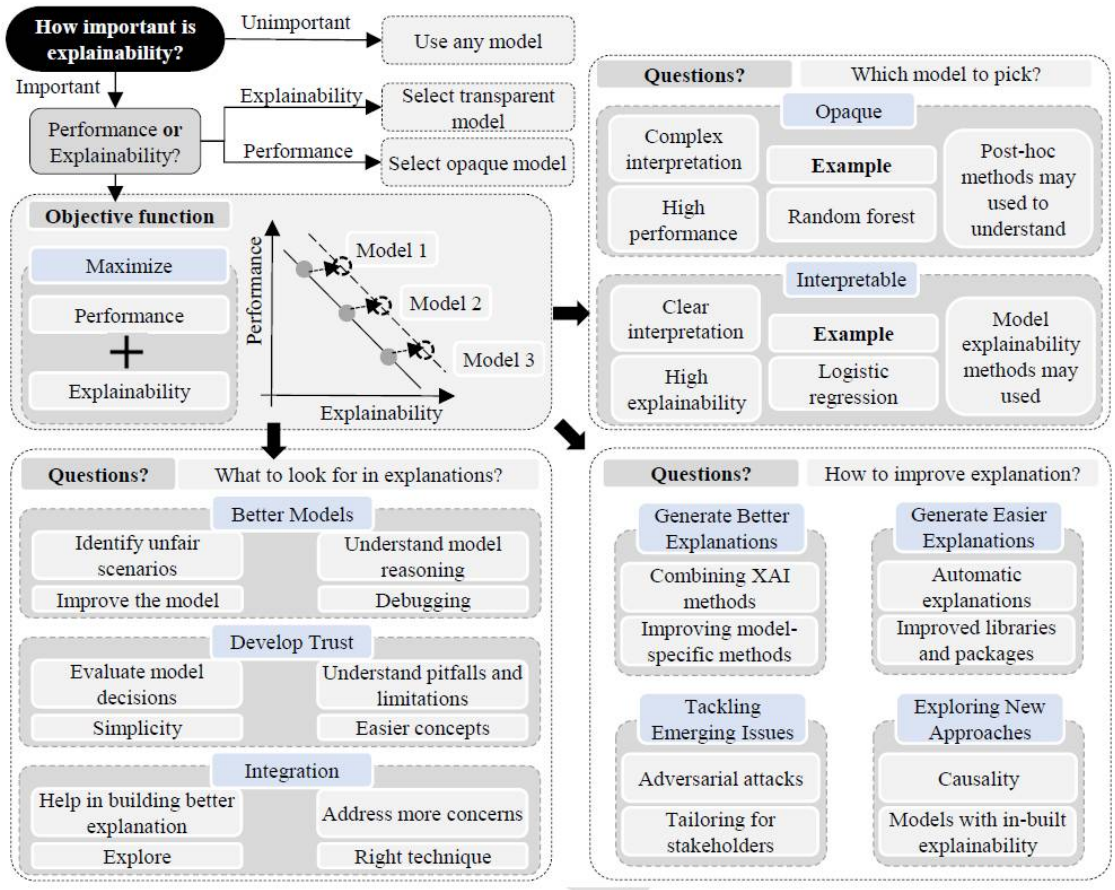
xAI.		2023-

Relationship among assessment methods for XAI and their desiderata

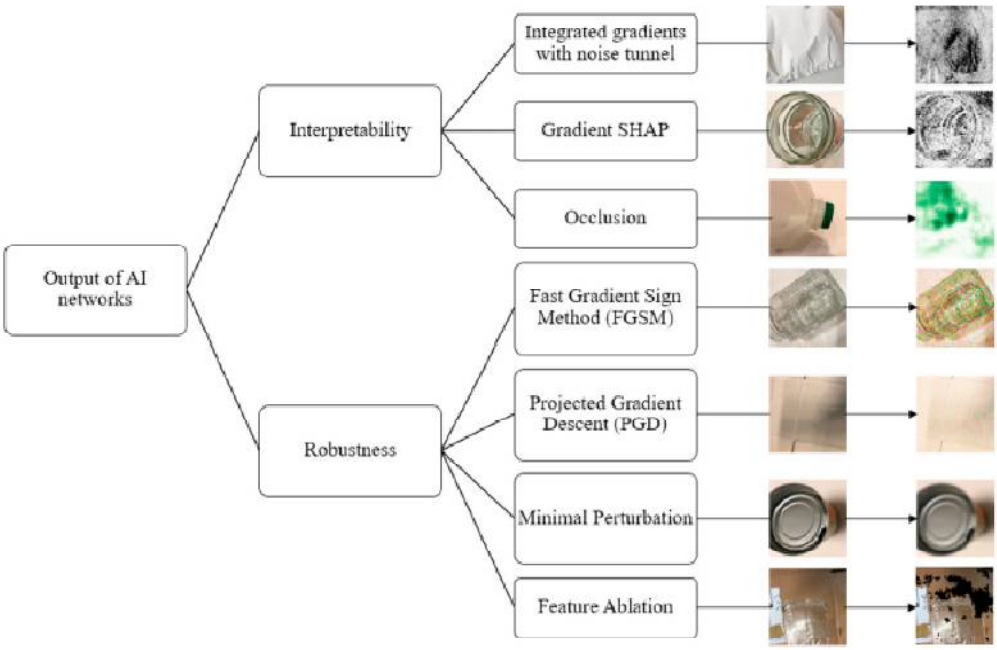


xAI.		2023-

Application of XAI



Interpretability and robustness of AI network outputs



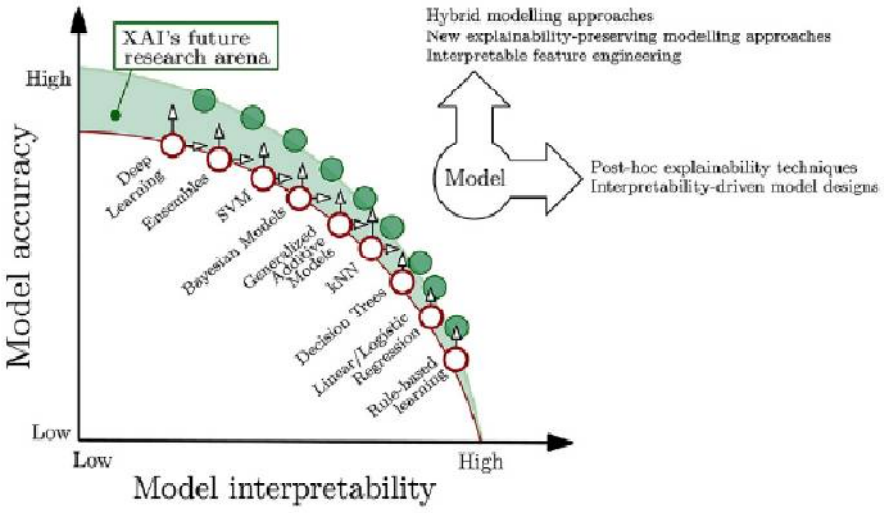
xAI.		2023-039

Trade-off between performance and explainability of AI models

- ✓ HBN: hierarchical Bayesian networks;
- ✓ SLR: simple linear regression;
- ✓ CRF: conditional random fields;
- ✓ MLN: Markov logic network;
- ✓ AOG: stochastic and–or graphs;
- ✓ XGB:extreme gradient boosting;
- ✓ GAN: generative adversarial network;

xAI.		2022-159

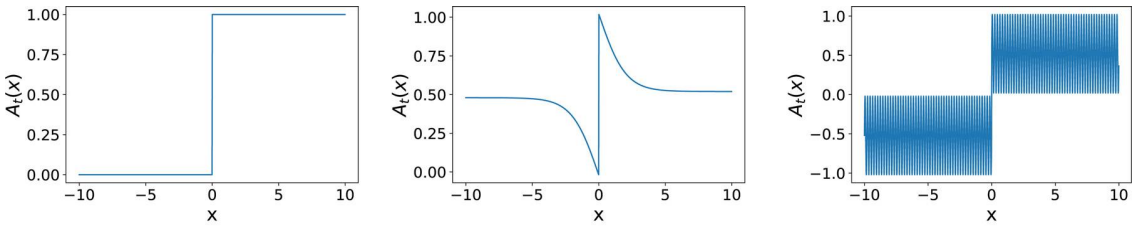
Interpretability of AI method



Transfer functions (TFs)

xAI. 2022-058

x

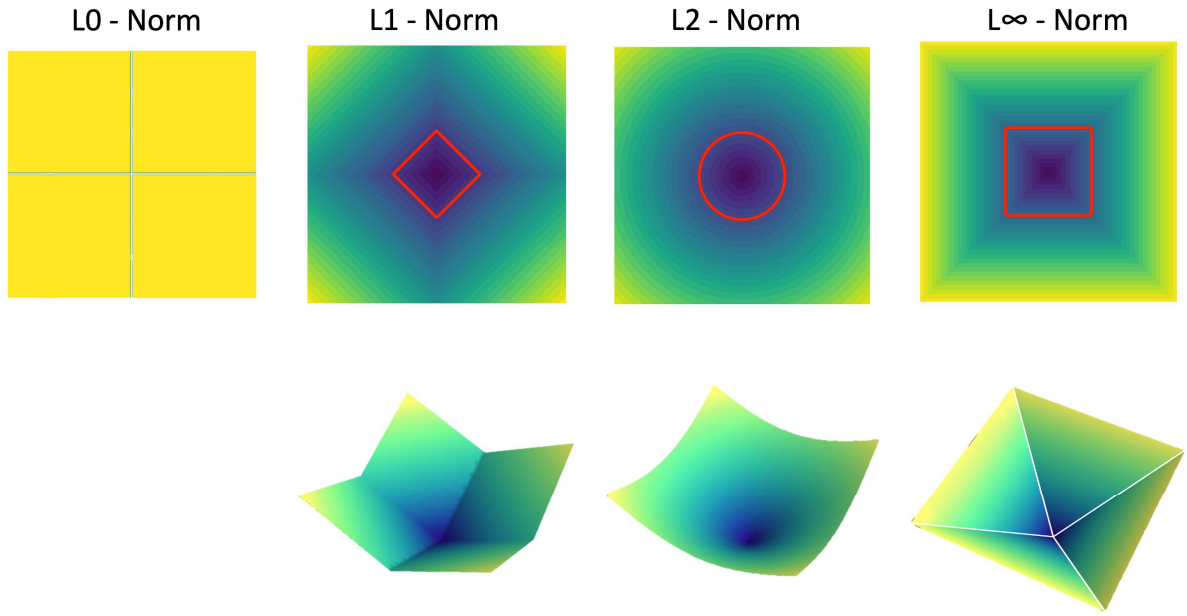


(a) Hard Sigmoid Activation (b) Inverse Sigmoid Activation (c) Sinusoidal Sigmoid Activation

L0, L1, L2, Linf Norms

xAI. 2022-

Graphical visualisation of different Lp-norms



- ✓ L0-norms (which is not really a norm by definition)
- ✓ L1-norm (also known as Manhattan distance)
- ✓ L2-norm (known as the Euclidean distance)

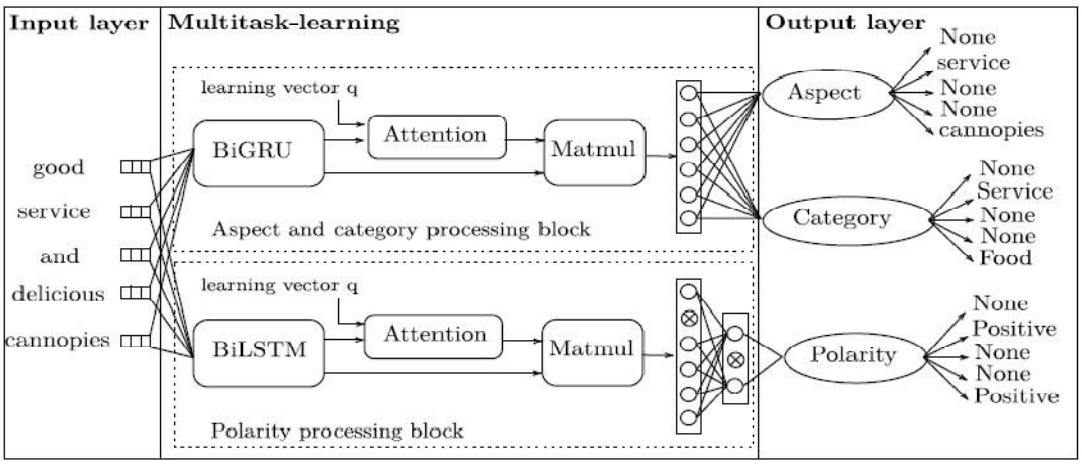
Architectures

Computational Nets (CNs)

xAI. 2023-144

Architecture of ASAM

Attention based Sentiment Analysis Method (ASAM)



Models

xAI. 2023-158

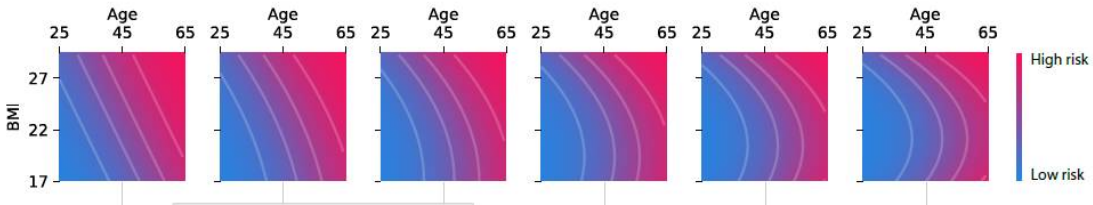
Gradient boosted tree models

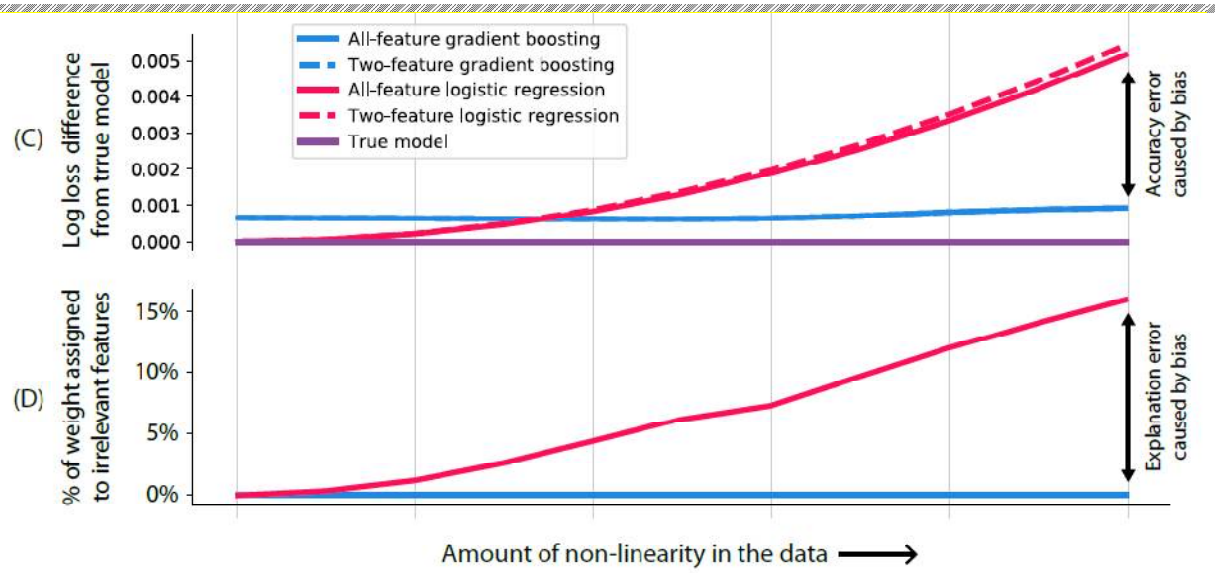
+ More accurate than neural networks
+ More interpretable than linear models

(A)

	Gradient Boosted Trees	Linear Model	Neural Network
NHANES I Mortality (C-statistic)	**0.821**	0.813	0.816
CRIC Kidney Disease (area under the PR curve)	**0.890**	0.871	0.872
Hospital Procedure Duration (R^2 value)	**0.674**	0.595	0.629

(B)





! (A) Gradient boosted tree models outperform both linear models and neural networks on medical datasets

! (B-D) Linear models exhibit explanation error as well as accuracy error inthe presence of non-linearity

! (B) Data generating models used for the simulation, ranging from linear to quadratic along the body mass index (BMI) dimension

! (C) The test performance of linear logistic regression (red) is better than gradient boosting (blue) up until a specific amount of non-linearity. Not surprisingly,the bias of the linear model is higher than the gradient boosting model as shown by the steeper slope with increase the non-linearity.

! (D) As the true function becomes more non-linear the linear model assigns morecredit (coefficient weight) to features that were not used by the data generating model.

xAI. 2023-144

Explanatory texts by ECDM-SDAM methodology to justify its final ranking

Positive explanatory text

- The user should book at the restaurant *The Ivy* since it obtains the highest overall rating.
- Its criterion of greatest interest, *food*, reaches a rating of 8.9 out of 10.
- The pie, crab, steak tartare, and liver stand out positively.
- Two of the expert sentences that most benefit this restaurant being selected as the best are: *"good service and delicious food"* and *"constantly the best."*

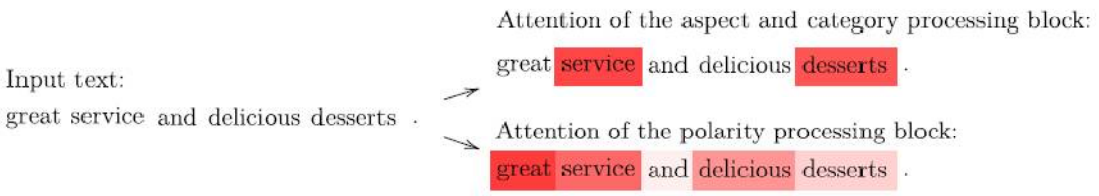
Negative explanatory text

- The restaurant *The Oxo Tower* is high quality although it is the last one of the ranking, so we identify its weakest points.
- Its most detrimental criterion is *drinks*.
- The acoustics, manager, and waiter stand out negatively.
- Two of the expert sentences that most harm to this restaurant being selected as the last one are: *"poor service, meagre portions"* and *"not worth the trip"*.

✓ SDAM: Subgroup Discovery and Attention Mechanisms

✓ ECDM: Explainable Crowd Decision Making

Visualization of weights of two attention layers of ASAM model



✓ Color intensity indicates relevancy or attention value of the word

Weighted Collective Evaluation matrix with identified criteria from the explanations

	restaurant	food	service	drinks	ambience	location
Oxo Tower	0.27	0.28	0.181	0.0283	0.075	0.02837
J. Sheekey	0.28	0.31	0.187	0.028	0.0776	0.028
The Wolseley	0.274	0.29	0.186	0.028	0.071	0.028
The Ivy	0.283	0.3	0.19	0.029	0.077	0.028

BestCriterion

WorstCriterion

Significant rules extracted from the BOC tables using the Apriori-SD algorithm.

Restaurant	Rule	NWRAcc	Support	Confidence
The Ivy	*{meal}* → *positive*	0.52	0.027	1
The Ivy	*{pie}* → *positive*	0.51	0.022	1
The Ivy	*{crab}* → *positive*	0.51	0.022	1
The Ivy	*{menu}* → *positive*	0.51	0.022	1
The Ivy	*{steak tartare}* → *positive*	0.508	0.16	1
The Ivy	*{liver}* → *positive*	0.508	0.16	1
The Oxo Tower	*{drinks}* → *negative*	1	0.036	1
The Oxo Tower	*{acoustics}* → *negative*	0.75	0.023	1
The Oxo Tower	*{manager}* → *negative*	0.58	0.018	1
The Oxo Tower	*{waiter}* → *negative*	0.58	0.018	1

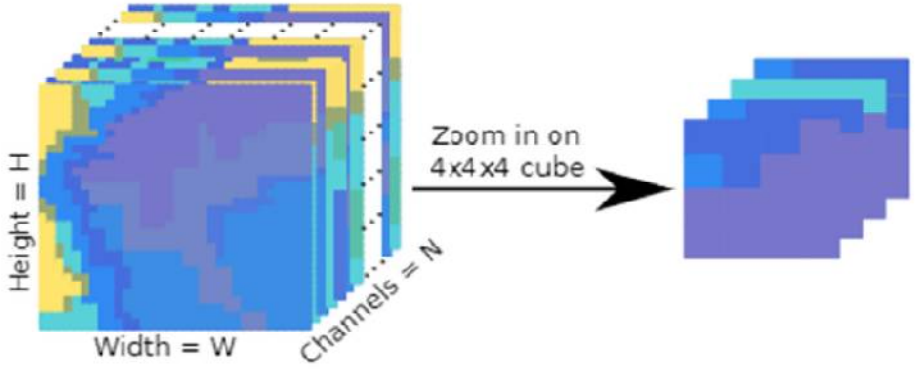
Significant top 3 positive and negative sentences.

Ranking 1 Positive sentences for The Ivy (food criterion)	1. Good service and delicious food. 2. Food and service as always terrific. 3. The food, service and sense of occasion was truly perfect.
Ranking 2 Positive sentences for The Ivy (restaurant criterion)	1. Constantly the best 2. Always very good 3. Will definitely revisit for a special occasion.
Ranking 3 Negative sentences for The Oxo Tower (any criterion)	1. Not worth the trip 2. Poor service, meagre portions 3. Overall, therefore, it is poor value and plays to the tourist market.

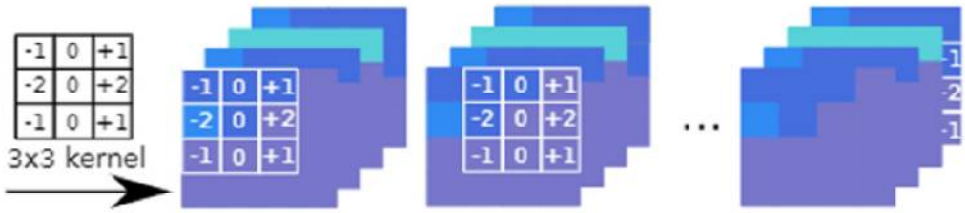
Convolution-
2D-3D

xAI.		2022-

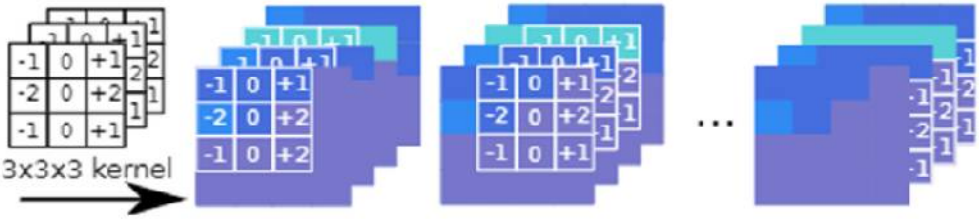
2D and 3D convolution on 3D data cube to generate feature maps



(a) Example of a 3d raster, or data cube.



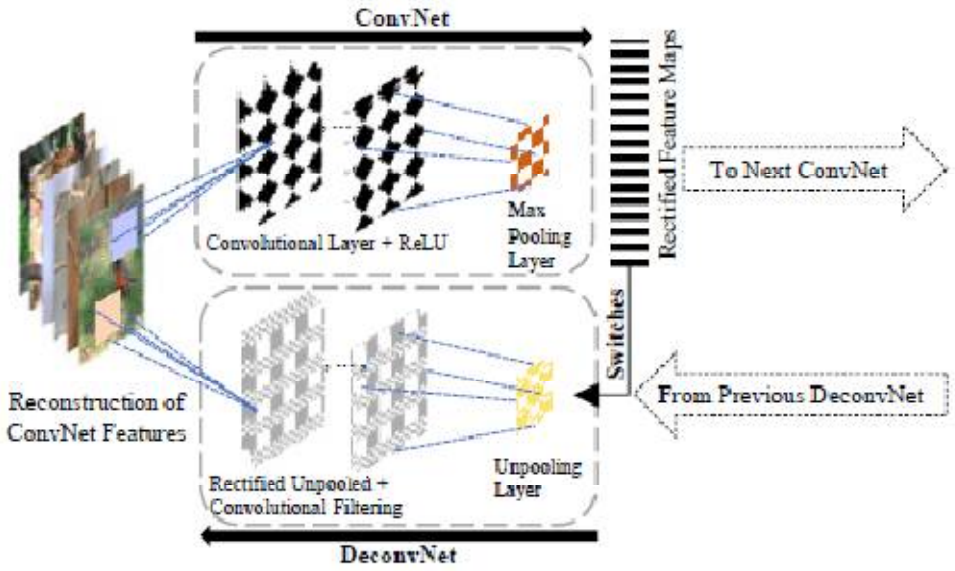
(b) Example 2D convolution to generate spatial-wise features.



(c) Example 3D convolution to generate spatial-channel-wise features.

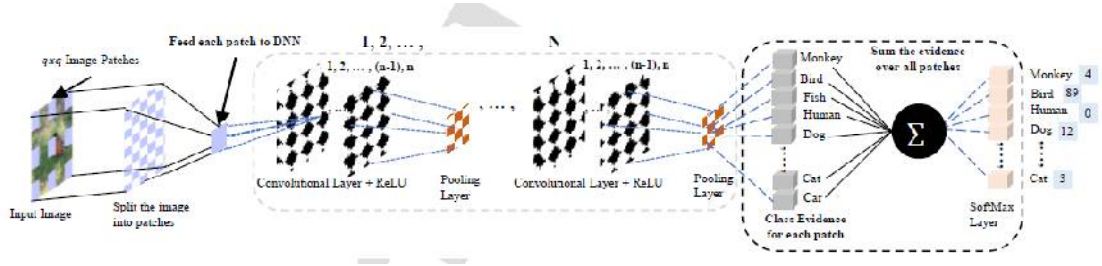
xAI.		2023-30

DeconvNet

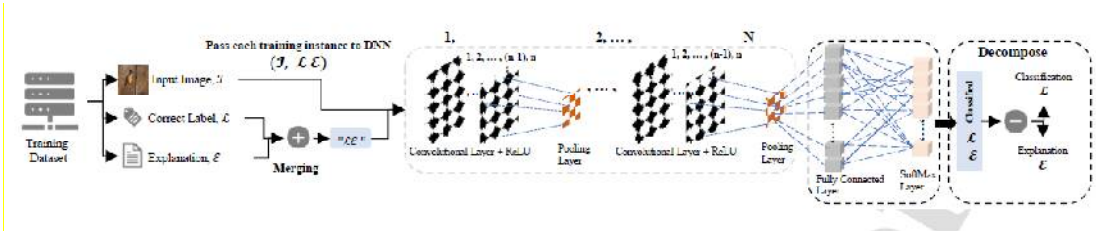


- ✓ Every layer in the ConvNet has a DeconvNet linked to it. It allows a continuous route back to the original input
- ✓ ConvNet receives an image and computes features across all layers.

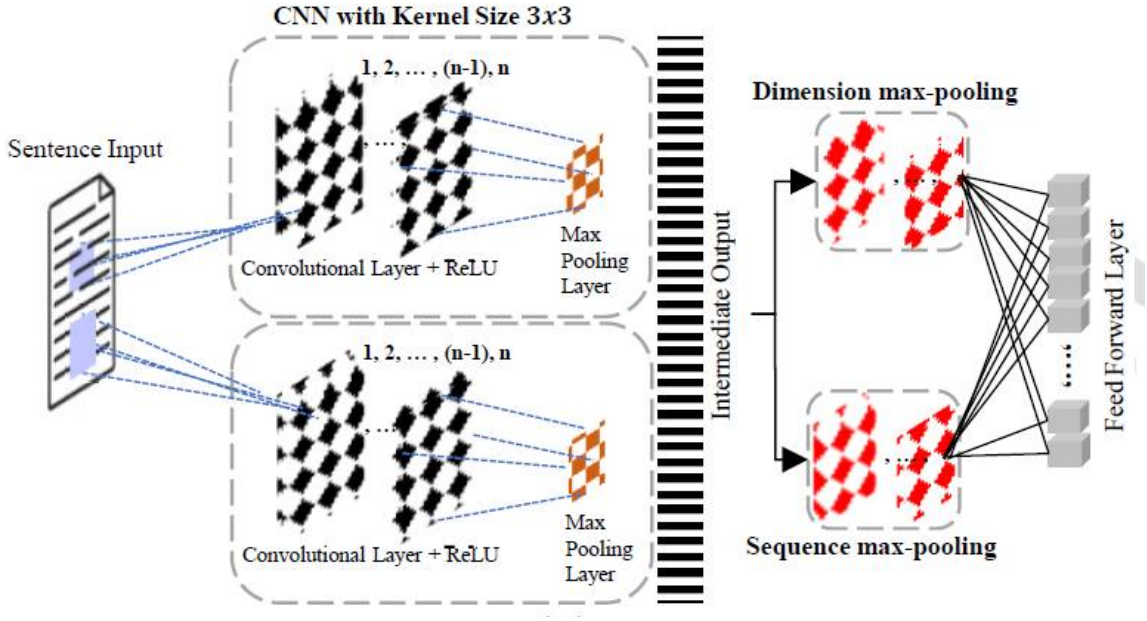
BagNets

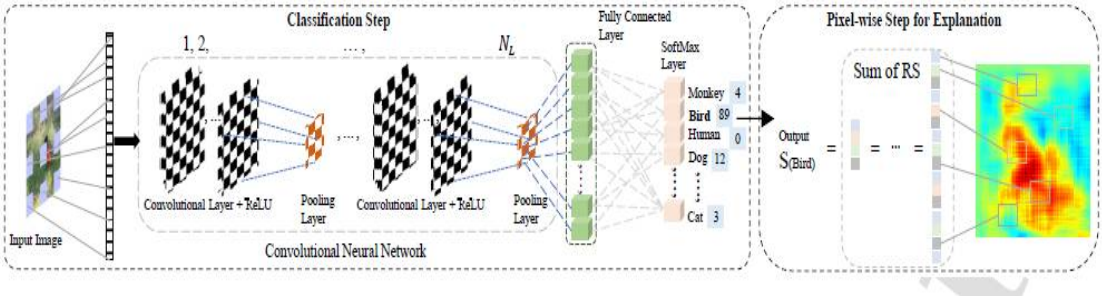


- o Input is first split into q × q patches.
- o Each patch is passed to the DNN to extract the evidence score
- o Taken sum of the class evidence scores overall patches →
- o Reaches final image classification decision

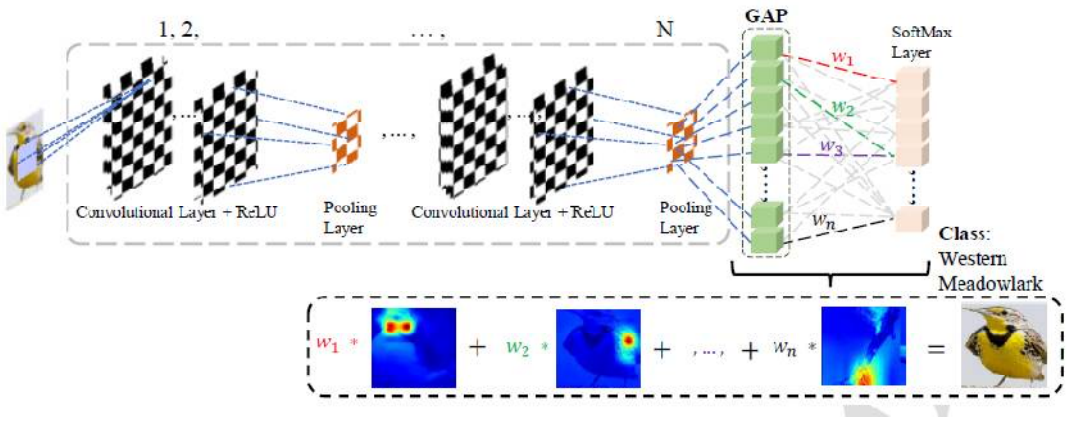


Deep Taylor Decomposition





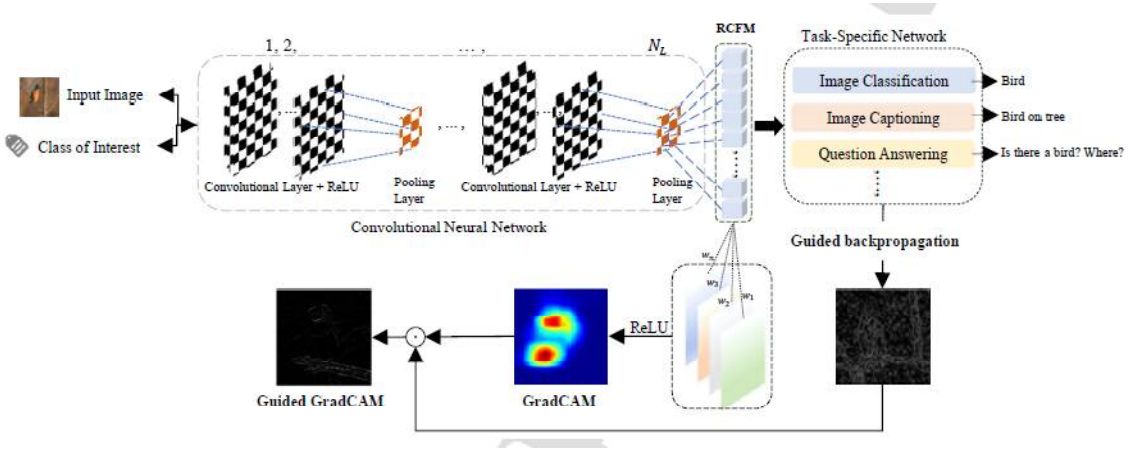
- I step: input image has been identified as a bird
- II step: model's reception of the features is shown as a heat map based on the relevance scores estimated from each hidden layer
- Pixels surrounding the bird's location had a substantial impact on the outcome
- Red regions: proved useful in the decision
- Blue regions: not helpful in decision



.

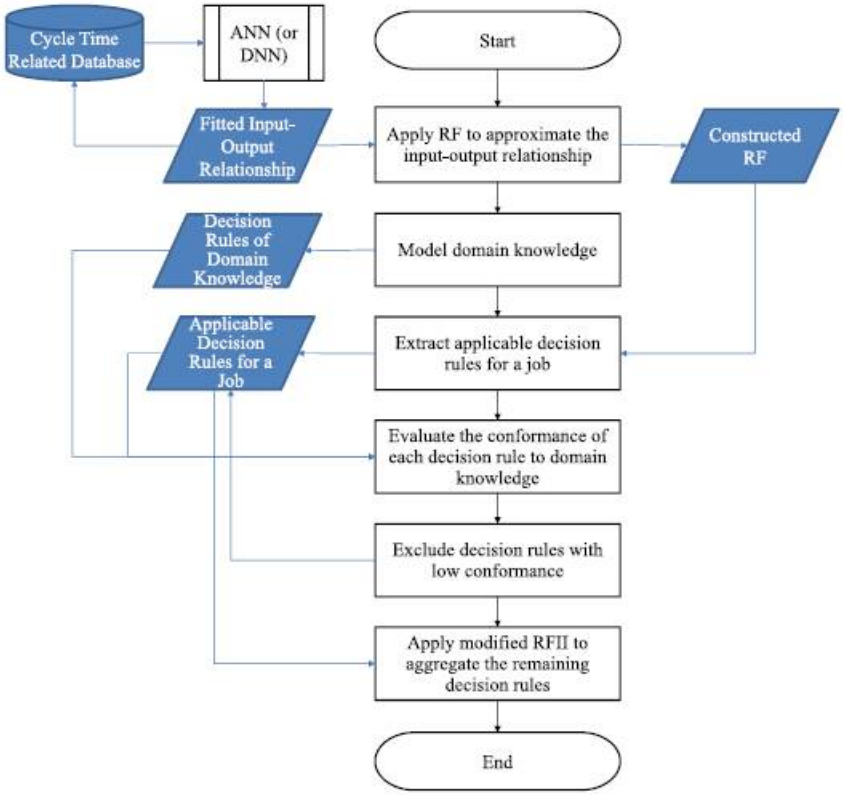
✓ Discriminative areas, distinct to each class, are highlighted in the CAM

DNN



xAI.	Cycle time prediction	2023-005

Process of methodology



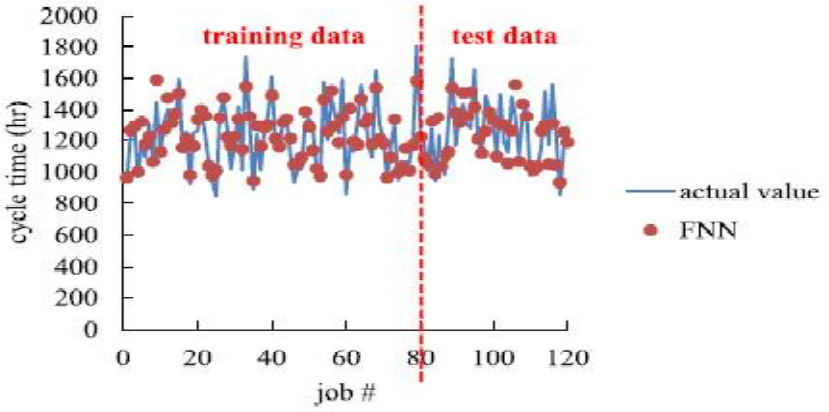
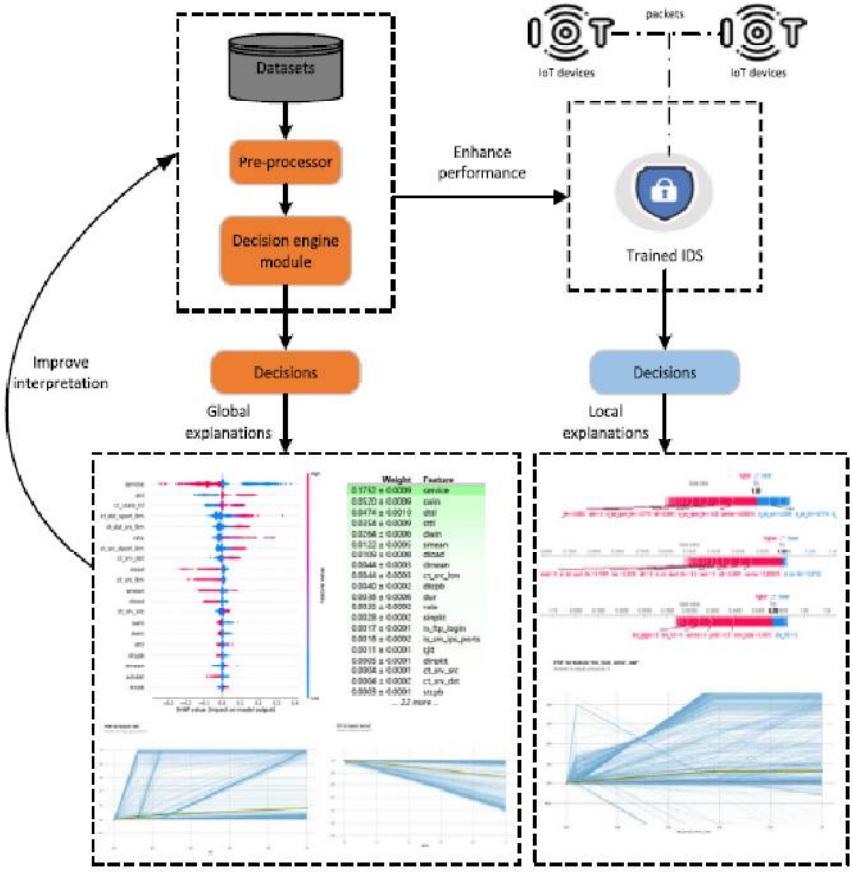


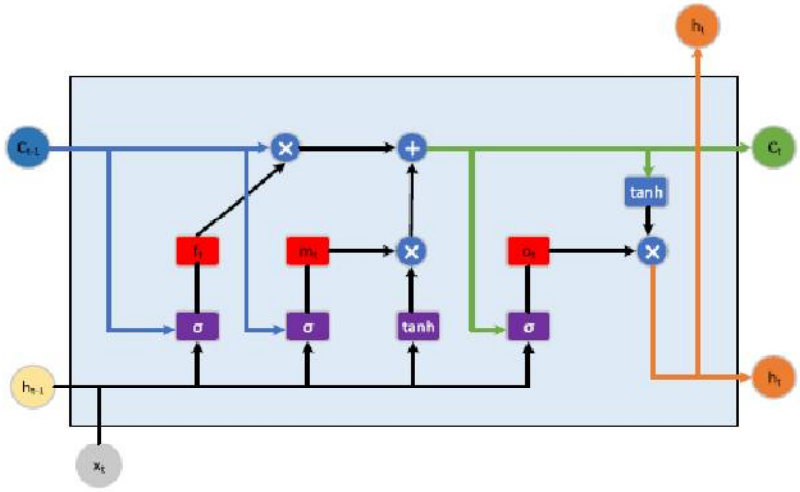
Fig. 2. Prediction results.

xAI.	S: Shapley Additive exPlanations P: Permutation Feature Importance I: Individual Conditional Expectation P: Partial Dependence Plot	2023-057

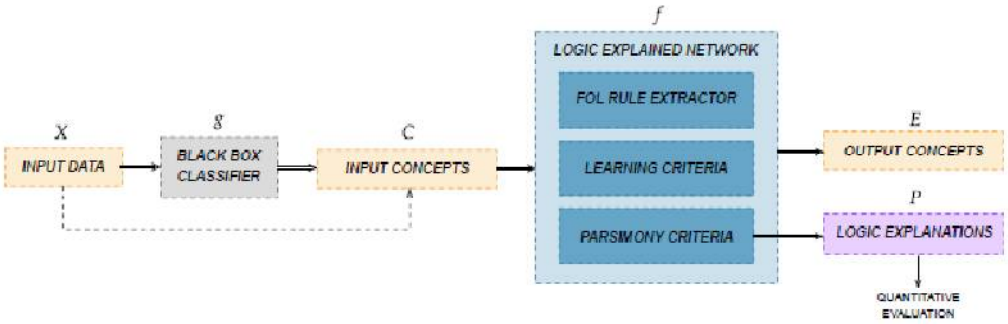
Overview of the structure of SPIP framework

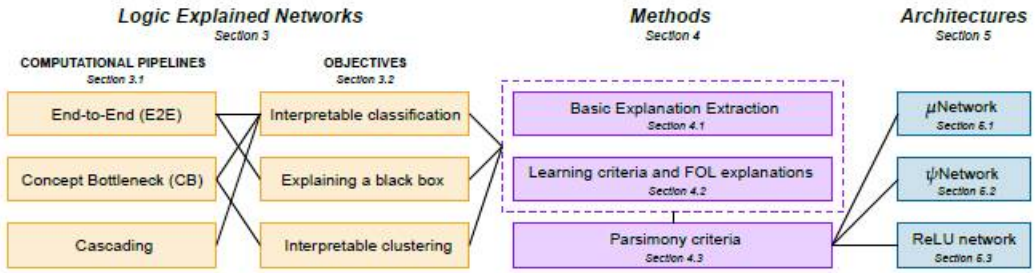


Long Short-Term Memory (RecNN)

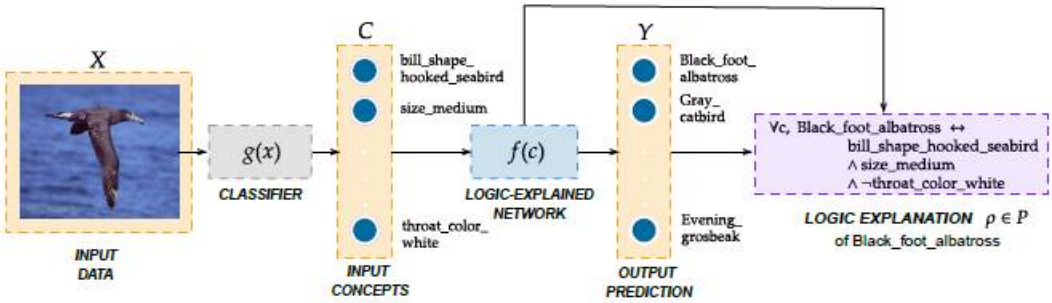


Logic Explained Networks (LENs, f)



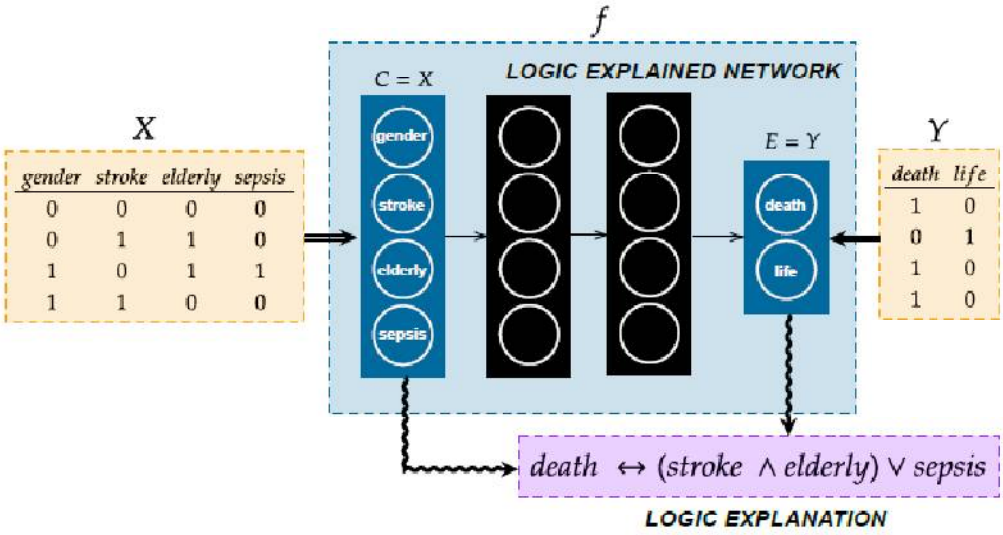


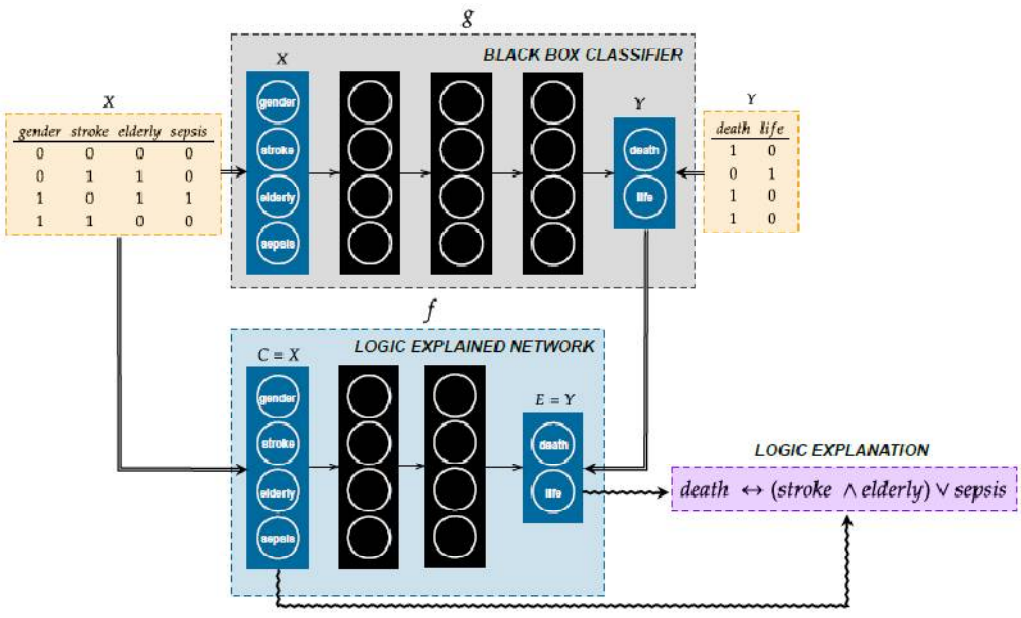
✓ Starting from the nodes on the left, each path that ends to one of the nodes on the right → creating a specific instance of the LEN framework



xAI.	Medical	2023-081

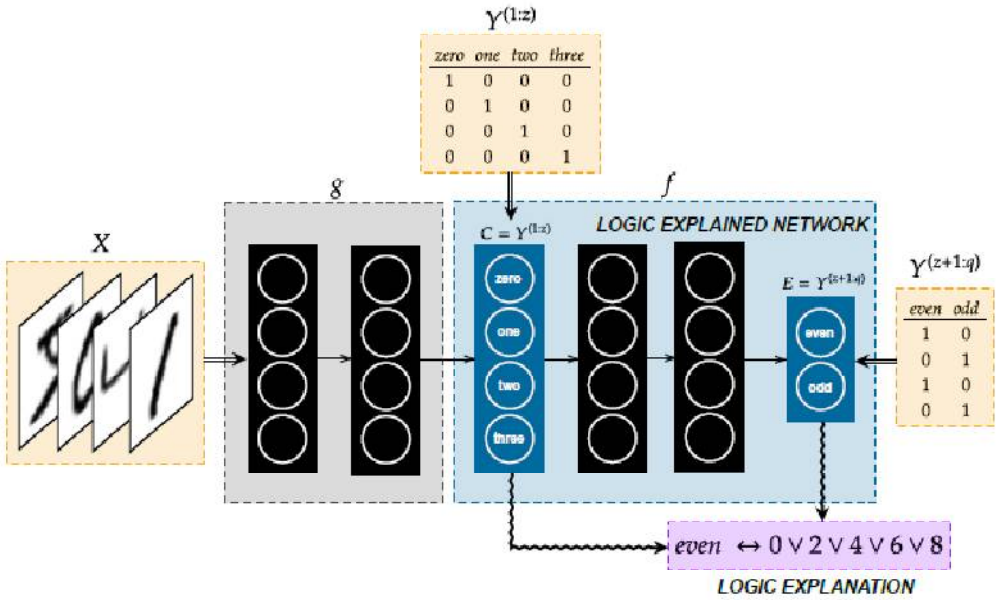
LEN (interpretable classification)
solves classification problem

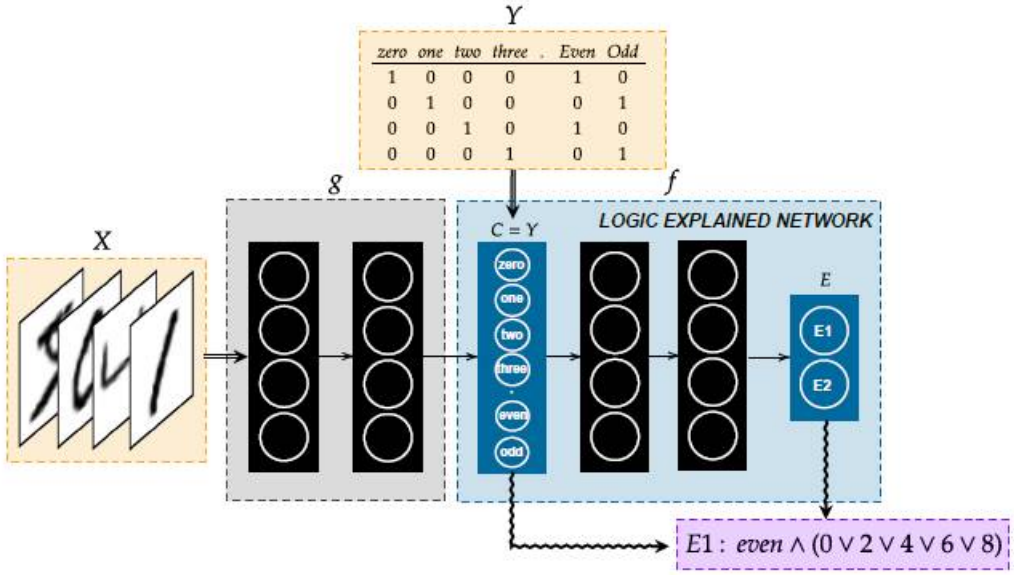


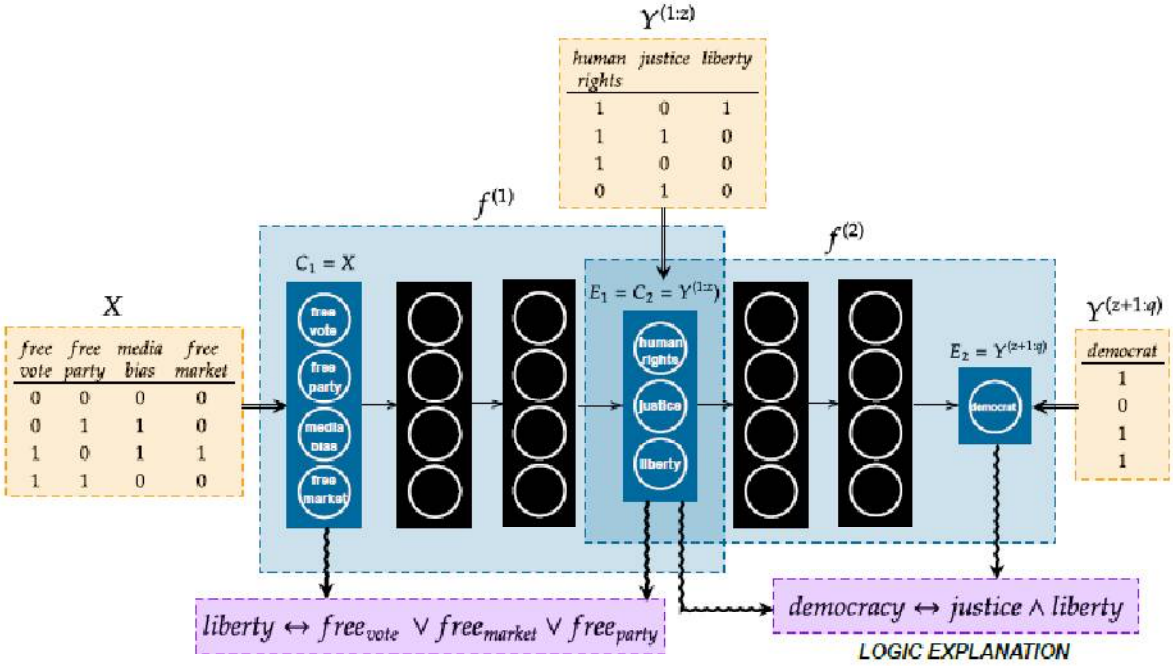


\+ LEN provides explanations of a black-box classifier

Cascading LENs

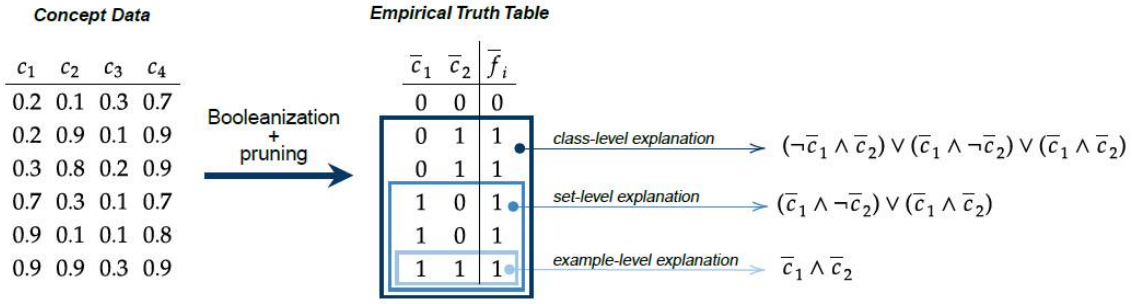




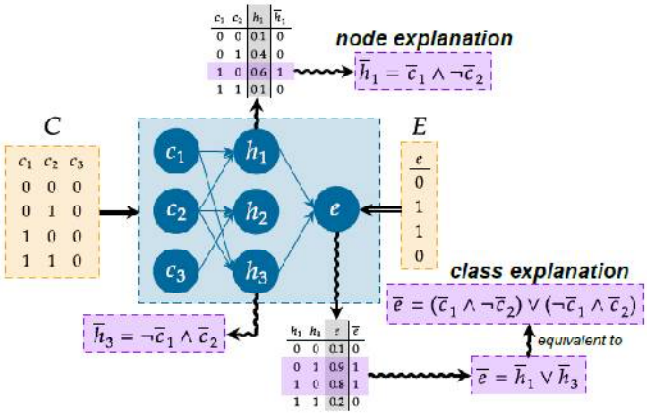


+ Cascading LENs provide explanations at different levels of granularity, implementing multiple interpretable classifications

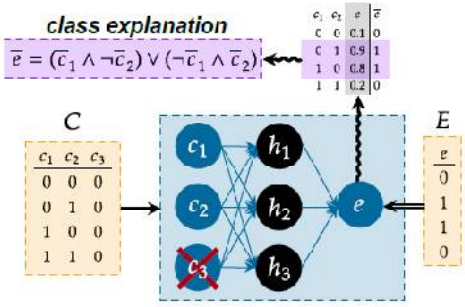
Empirical truth table T i
of i-th LEN output fi, with k = 4 and mi = 2



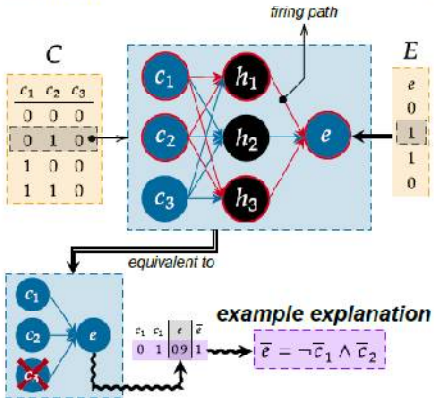
Out-of-the-box LENs



(a) ψ network (Section 5.1).

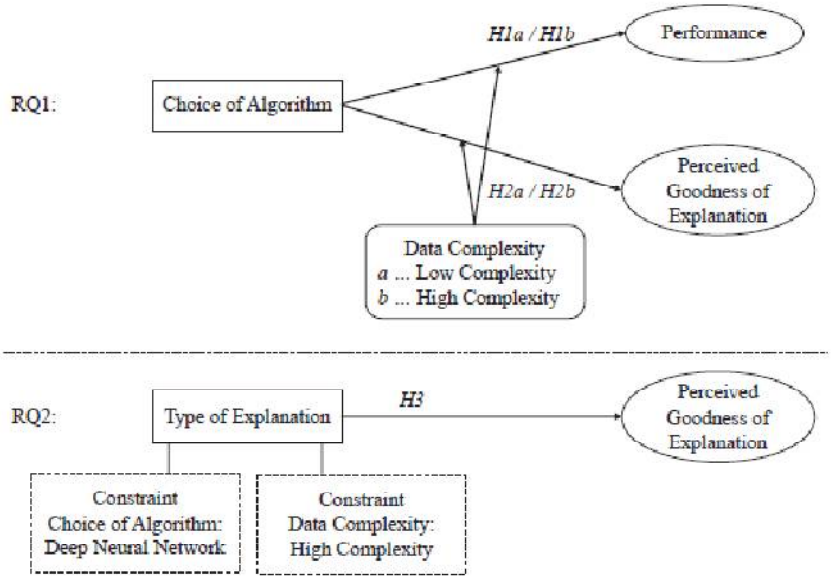


(b) μ network (Section 5.2).



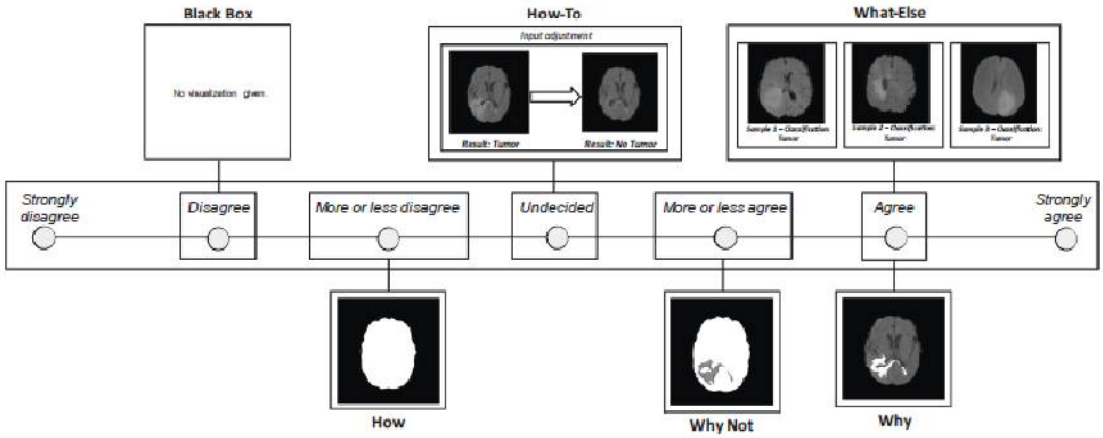
(c) ReLU network (Section 5.3).

	Model	Sample Rule
MIMIC-II	μ net	Death $\leftrightarrow$ stroke $\wedge$ age_HIGH $\wedge$ −atrial_fibrillation
	ReLU net	Death $\leftrightarrow$ stroke $\wedge$ age_HIGH $\wedge$ −atrial_fibrillation $\wedge$ ¬sapsi_first_LOW $\wedge$ ¬sapsi_first_HIGH
	ψ net	Death $\leftrightarrow$ stroke $\vee$ bun_first_NORMAL $\vee$ sapsi_first_HIGH
	Tree	Death $\leftrightarrow$ *formula is too long
	BRL	Death $\leftrightarrow$ (stroke $\wedge$ resp $\wedge$ age_LOW $\wedge$ sapsi_first_LOW) $\vee$ (stroke $\wedge$ age_HIGH $\wedge$ ¬age_LOW $\wedge$ ¬sapsi_first_LOW)
MNIST E/O	μ net	Even $\leftrightarrow$ One $\wedge$ Three $\wedge$ Five $\wedge$ Seven $\wedge$ Nine
	ReLU net	Even $\leftrightarrow$ (Zero $\wedge$ ¬One $\wedge$ ¬Two $\wedge$ ¬Three $\wedge$ ¬Four $\wedge$ ¬Five $\wedge$ ¬Six $\wedge$ ¬Seven $\wedge$ ¬Eight $\wedge$ ¬Nine) $\vee$ (Two $\wedge$ ¬Zero $\wedge$ ¬One $\wedge$ ¬Three $\wedge$ ¬Four $\wedge$ ¬Five $\wedge$ ¬Six $\wedge$ ¬Seven $\wedge$ ¬Eight $\wedge$ ¬Nine) $\vee$ (Four $\wedge$ Zero $\wedge$ One $\wedge$ Two $\wedge$ Three $\wedge$ Five $\wedge$ Six $\wedge$ Seven $\wedge$ Eight $\wedge$ Nine) $\vee$ (Six $\wedge$ ¬Zero $\wedge$ ¬One $\wedge$ ¬Two $\wedge$ ¬Three $\wedge$ ¬Four $\wedge$ ¬Five $\wedge$ ¬Seven $\wedge$ ¬Eight $\wedge$ ¬Nine) $\vee$ (Eight $\wedge$ ¬Zero $\wedge$ ¬One $\wedge$ ¬Two $\wedge$ ¬Three $\wedge$ ¬Four $\wedge$ ¬Five $\wedge$ ¬Six $\wedge$ ¬Seven $\wedge$ ¬Nine)
	ψ net	Even $\leftrightarrow$ (Six $\wedge$ Zero $\wedge$ One $\wedge$ Seven) $\vee$ (Six $\wedge$ Zero $\wedge$ One $\wedge$ Three) $\vee$ (Six $\wedge$ Zero $\wedge$ ¬Seven $\wedge$ ¬Three) $\vee$ (Six $\wedge$ ¬One $\wedge$ ¬Seven $\wedge$ ¬Three) $\vee$ (Zero $\wedge$ ¬One $\wedge$ ¬Seven $\wedge$ ¬Three) $\vee$ (Six $\wedge$ Zero $\wedge$ ¬One $\wedge$ ¬Seven $\wedge$ ¬Three)
	Tree	Even $\leftrightarrow$ *formula is too long
	BRL	Even $\leftrightarrow$ (Six $\wedge$ ¬Three $\wedge$ ¬(Five $\wedge$ ¬Two)) $\vee$ (Eight $\wedge$ ¬Nine $\wedge$ ¬Six $\wedge$ ¬Three $\wedge$ ¬(Five $\wedge$ ¬Two) $\vee$ (Four $\wedge$ ¬Nine $\wedge$ ¬Six $\wedge$ ¬Three $\wedge$ ¬(Eight $\wedge$ ¬Nine) $\wedge$ ¬(Five $\wedge$ ¬Two) $\wedge$ (Seven $\wedge$ Two)) $\vee$ (Two $\wedge$ One $\wedge$ Six $\wedge$ Three $\wedge$ (Eight $\wedge$ Nine) $\wedge$ ¬(Five $\wedge$ ¬Two) $\wedge$ ¬(Four $\wedge$ ¬Nine) $\wedge$ ¬(Seven $\wedge$ ¬Nine) $\wedge$ ¬(Seven $\wedge$ ¬Two)) $\vee$ (Four $\wedge$ ¬Six $\wedge$ ¬Three $\wedge$ ¬(Eight $\wedge$ ¬Nine) $\wedge$ ¬(Five $\wedge$ ¬Two) $\wedge$ ¬(Four $\wedge$ ¬Nine) $\wedge$ ¬(Seven $\wedge$ ¬Nine) $\wedge$ ¬(Seven $\wedge$ ¬Two) $\wedge$ ¬(Two $\wedge$ ¬One)) $\vee$ (Zero $\wedge$ ¬Four $\wedge$ ¬Nine $\wedge$ ¬Six $\wedge$ ¬Three $\wedge$ ¬(Eight $\wedge$ ¬Nine) $\wedge$ ¬(Five $\wedge$ ¬Two) $\wedge$ ¬(Four $\wedge$ ¬Nine) $\wedge$ ¬(Nine $\wedge$ ¬Zero) $\wedge$ ¬(One $\wedge$ ¬Two) $\wedge$ ¬(Seven $\wedge$ ¬Nine) $\wedge$ ¬(Seven $\wedge$ ¬Two) $\wedge$ ¬(Two $\wedge$ ¬One))



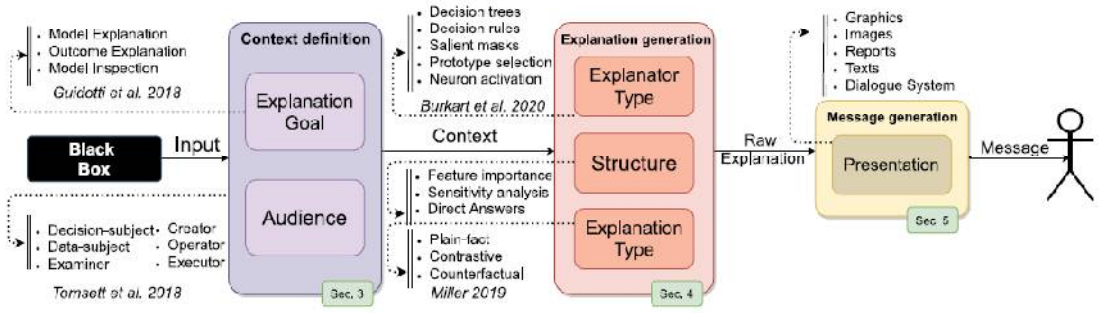
Research Question (RQ)

Average End User Ratings of Explanation Type
Visualized on Likert Scale

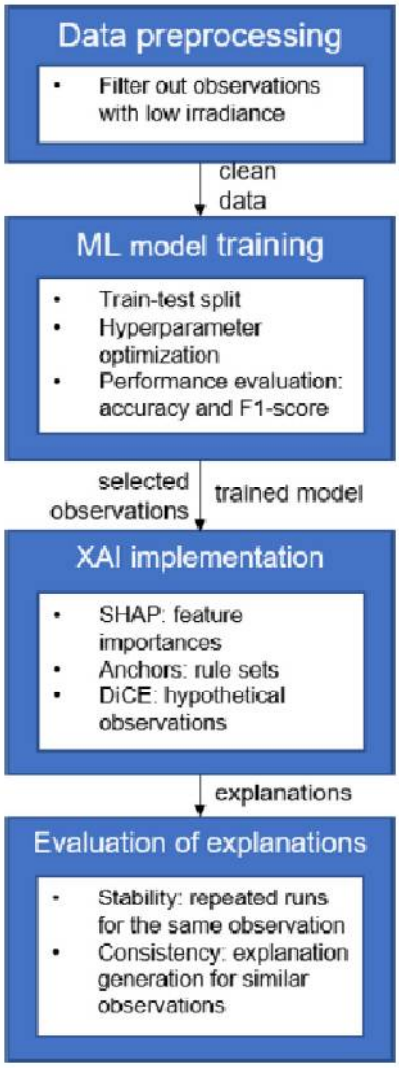


xAI.		2023-134

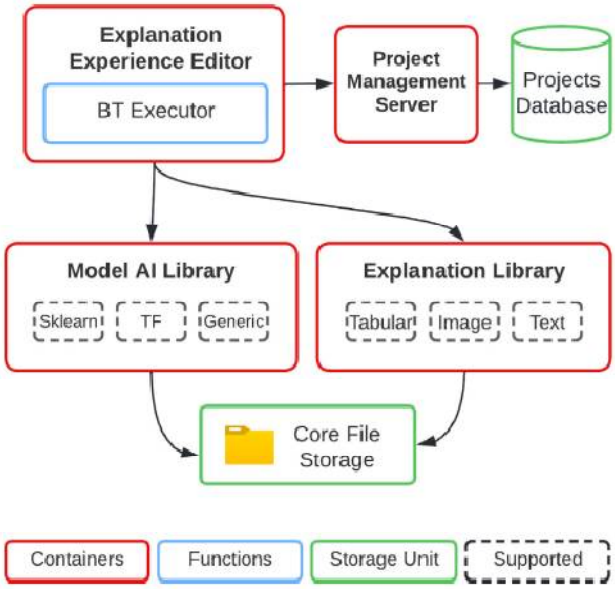
Roadmap for selecting XAI systems using
Natural language explanations (NLEx)

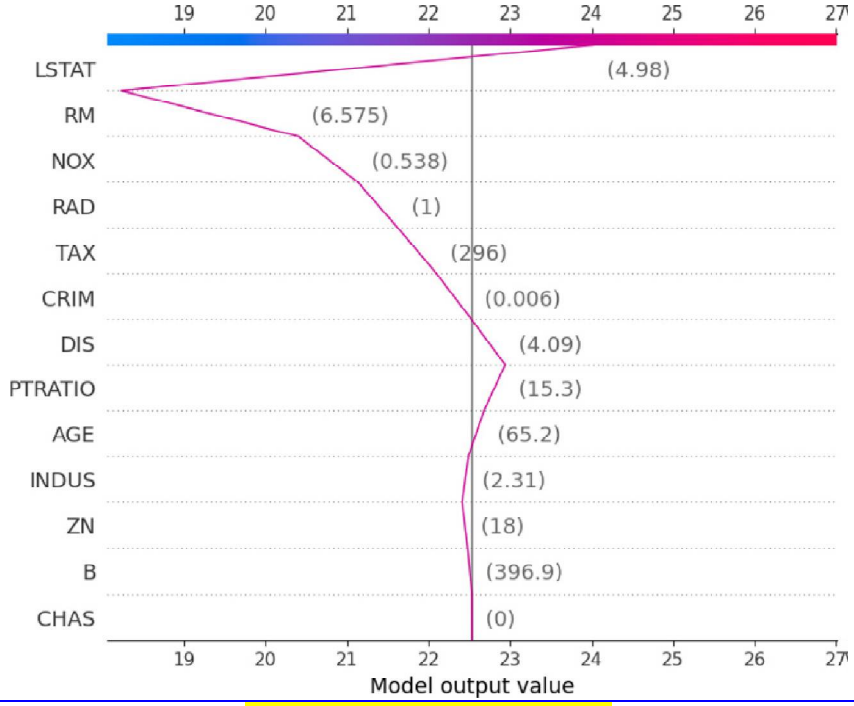


Method Flow xAI

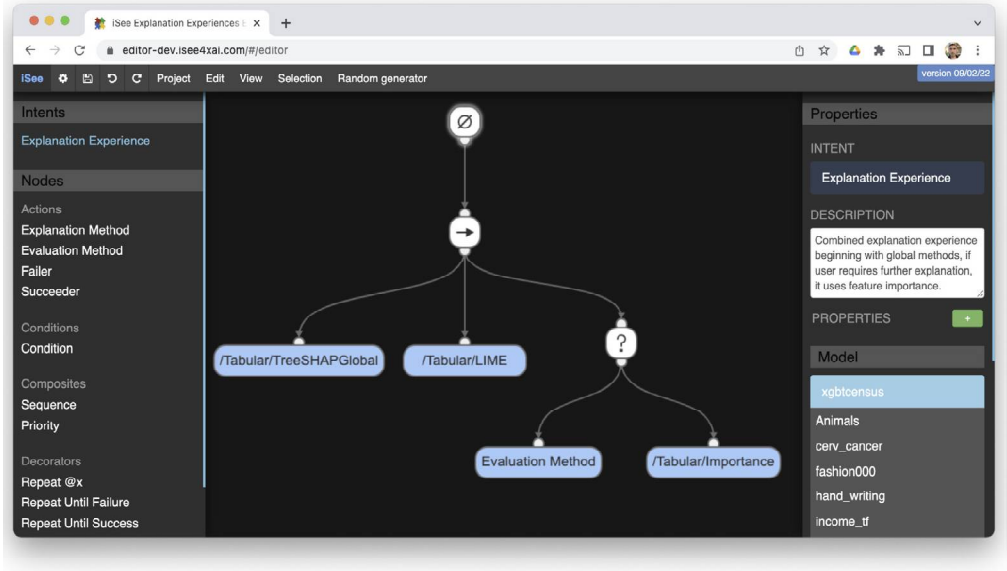


Software architecture Flow-diagram Evolving (Safe)





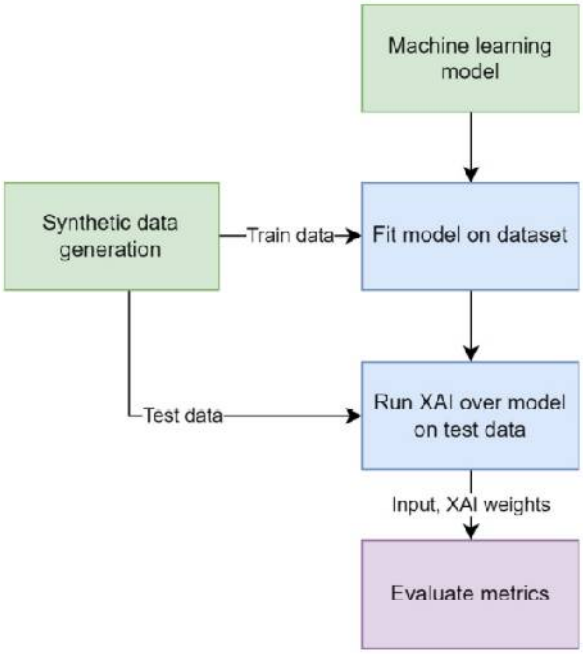
Model selection in E3 software



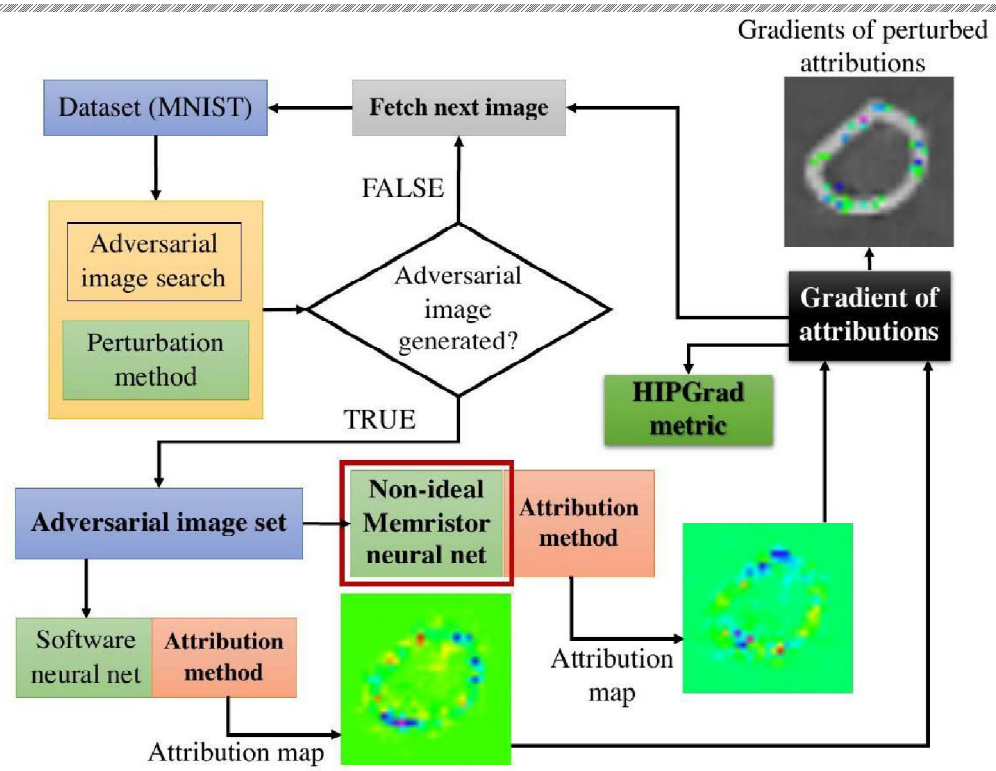
Snapshot of the E3 software

xAI.		2023-142

xAI-method



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



xAI.		2023-160

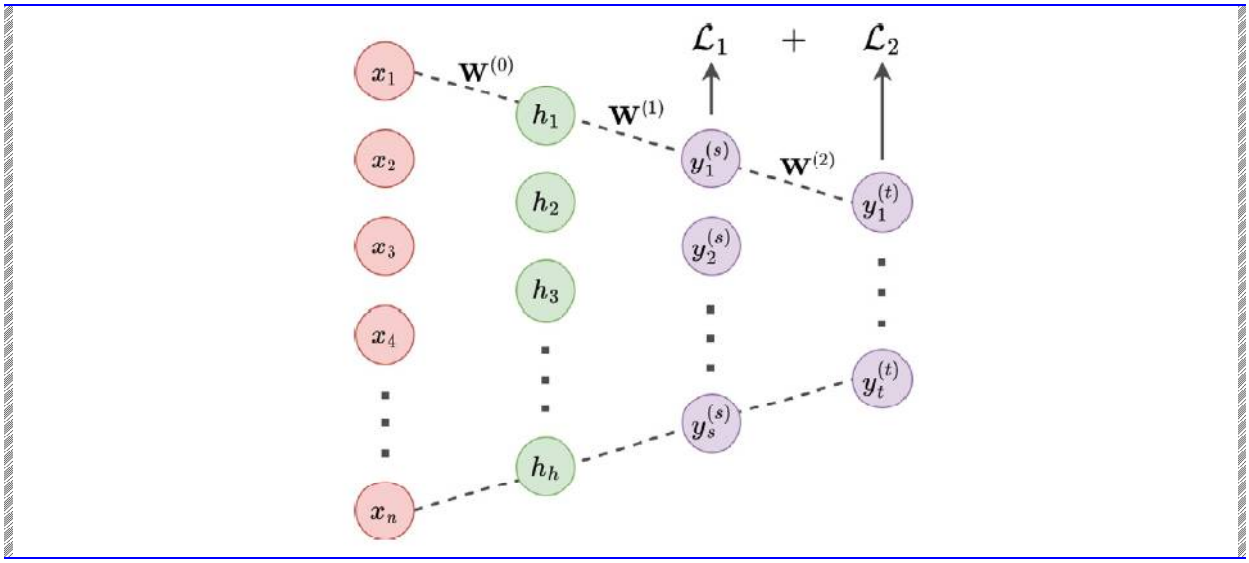
Libraries

Library	Faithfulness	Robustness	Localisation	Complexity	Axiomatic	Randomisation
`Captum` (2)	1	1	0	0	0	0
`AIX360` (2)	2	0	0	0	0	0
`TorchRay` (1)	0	0	1	0	0	0
`Quantus` (27)	**9**	**4**	**6**	**3**	**3**	**2**

AIX360 (Arya et al., 2019), captum (Kokhlikyan et al., 2020), TorchRay (Fong et al., 2019) and Quantus)

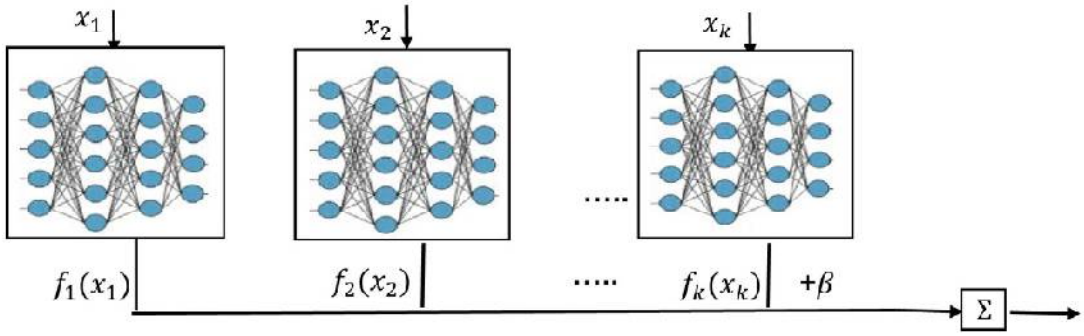
xAI.	PLENARY (exPlaining bLack-box modEls in Natural lAnguage thRough fuzzY linguistic summaries),	2022-016

Architecture of multi-task neural network



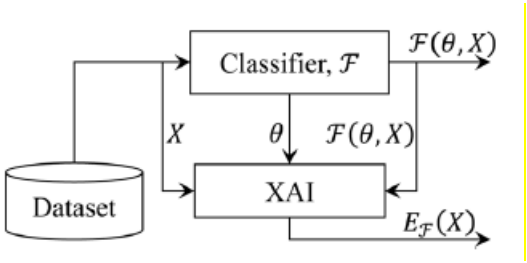
xAI. 2022-045

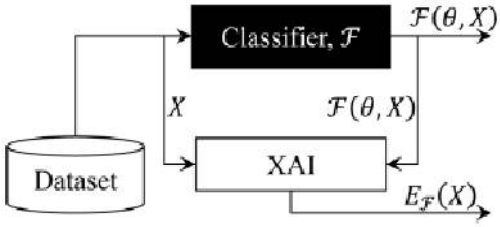
The Neural Additive Model for binary classification

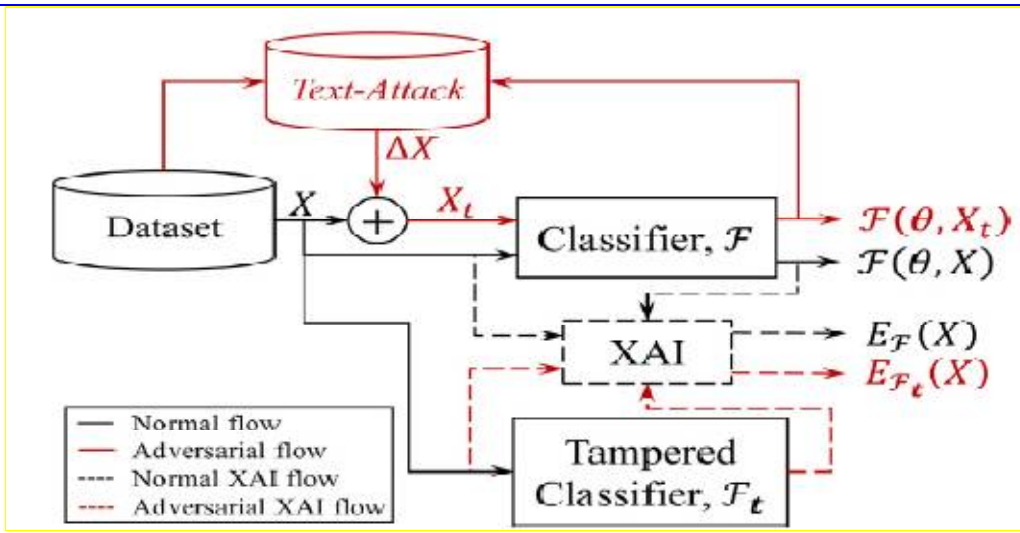


xAI. 2022-058

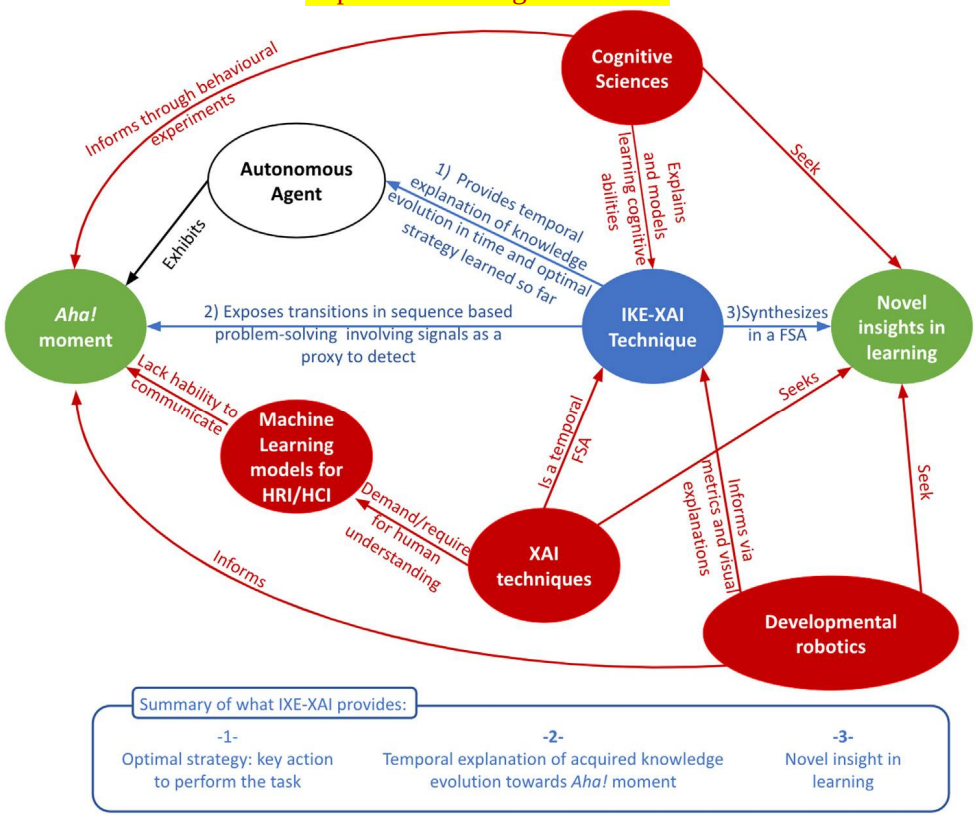
xAI







IKE-XAI method
Implicit Knowledge Extraction

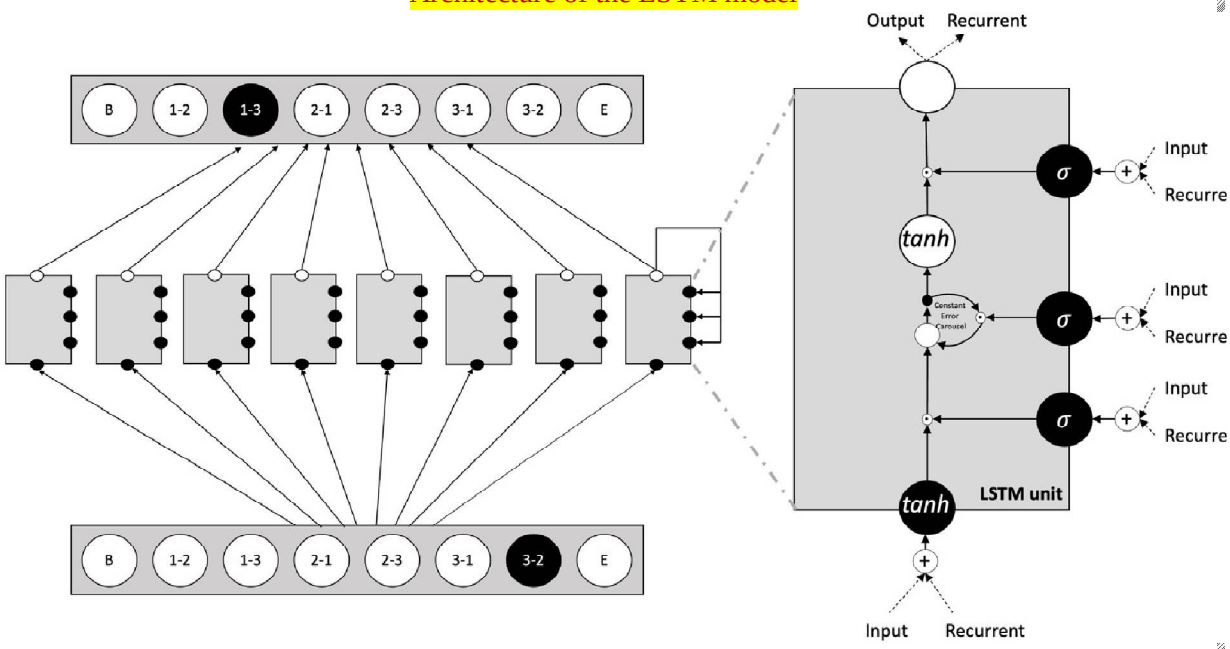


FIELDS of RESEARCH	Developmental Robotics : Machine learning	Cognitive developmental psychology : Piaget's concrete operational stage	EXplainable AI: Sequence Models Graphical Summary Representation.
USE CASE	Towers of Hanoï - Concrete Problem Solving - Means-ends heuristics (General Problem Solver - Newell & Simon, 1972)		
	- STEP 1 -	- STEP 2 -	- STEP 3 -
EXPERIMENTAL PROTOCOL	Reinforcement learning phase : Exploration and Training	Sequence learning phase: Implicit rule encoding	XAI phase : Understanding the AA learning process
	AA – QLearning Agent Recording of number of movements	Trained LSTM model* AA sequences of movements	Post-hoc Rule extraction process and graph visualisation techniques Data set of recorded hidden patterns
RAW DATA obtained	Extraction of AA sequences of movements	Data set of recorded hidden patterns	Graphs of AA behavior at different stage of training
RESULTS	**An AA trained to perform TOH** whose behavior is observable through its movements sequences	**A trained LSTM model having developed a latent representation of the TOH** including all possible actions and their resulting states	**Accurate and specific graphical representations** for a targeted audience = AI researchers (Dev. robotics community)

* Trained LSTM model : a RNN with LSTM units that learned sequences of moves generated from Towers of Hanoï **to predict the next move according the past and current ones**

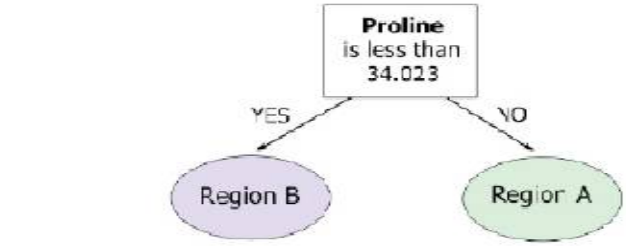
xAI.		2022-077

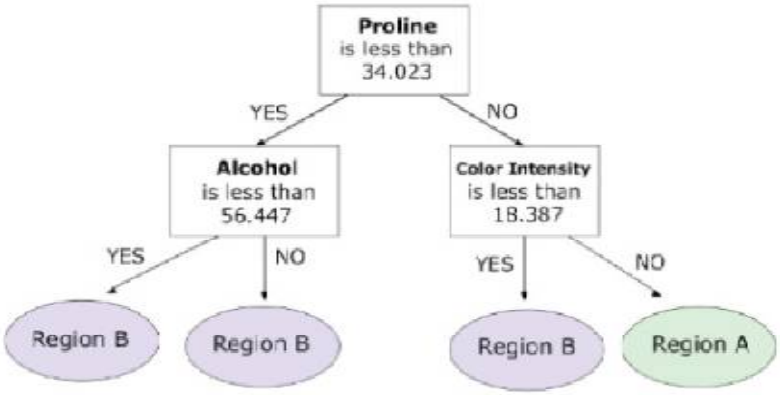
Architecture of the LSTM model



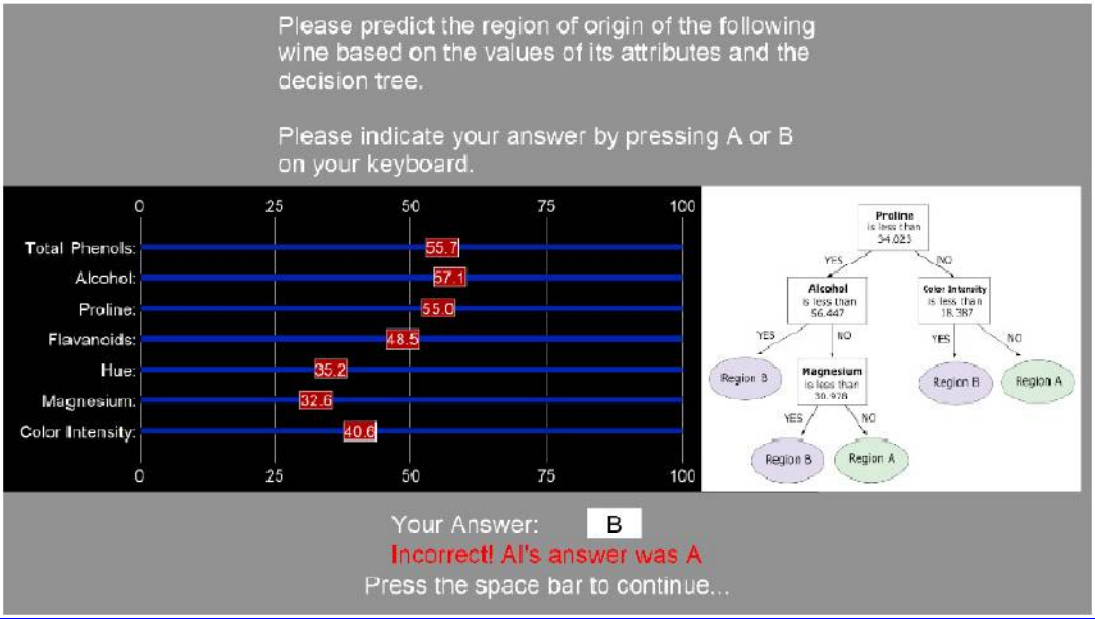
xAI.		2022-

Integrated Evaluation Framework--xAI



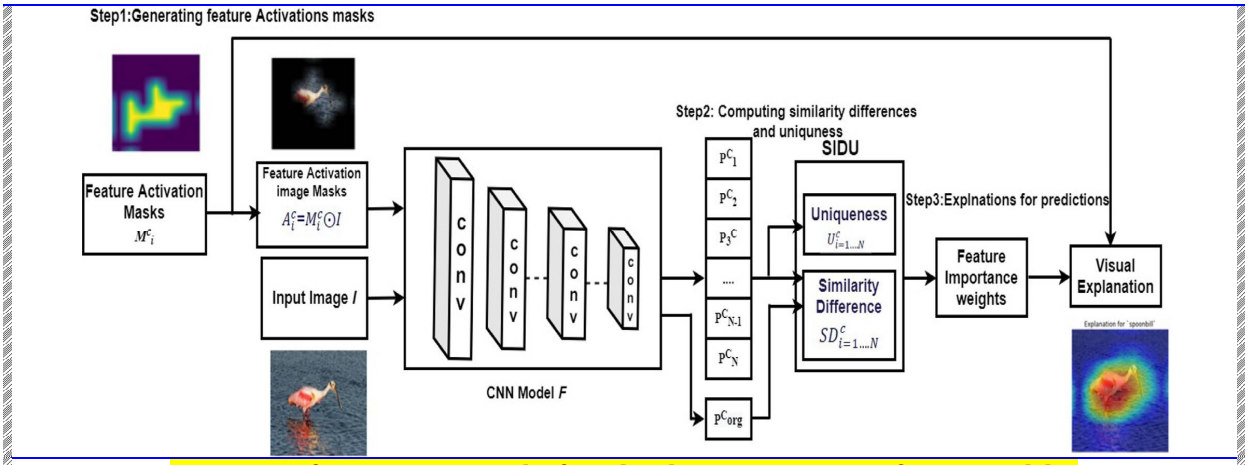


Decision trees in categorization tasks

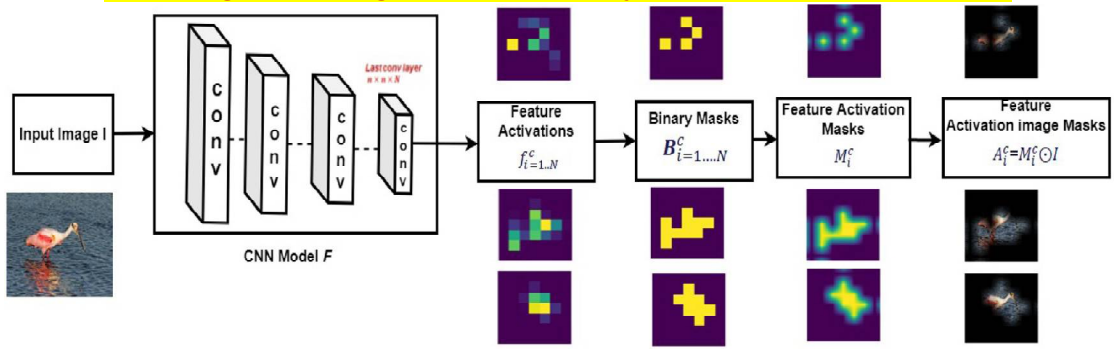


xAI.	Similarity Difference and Uniqueness (SIDU)	2022-124

Block diagram of SIDU.

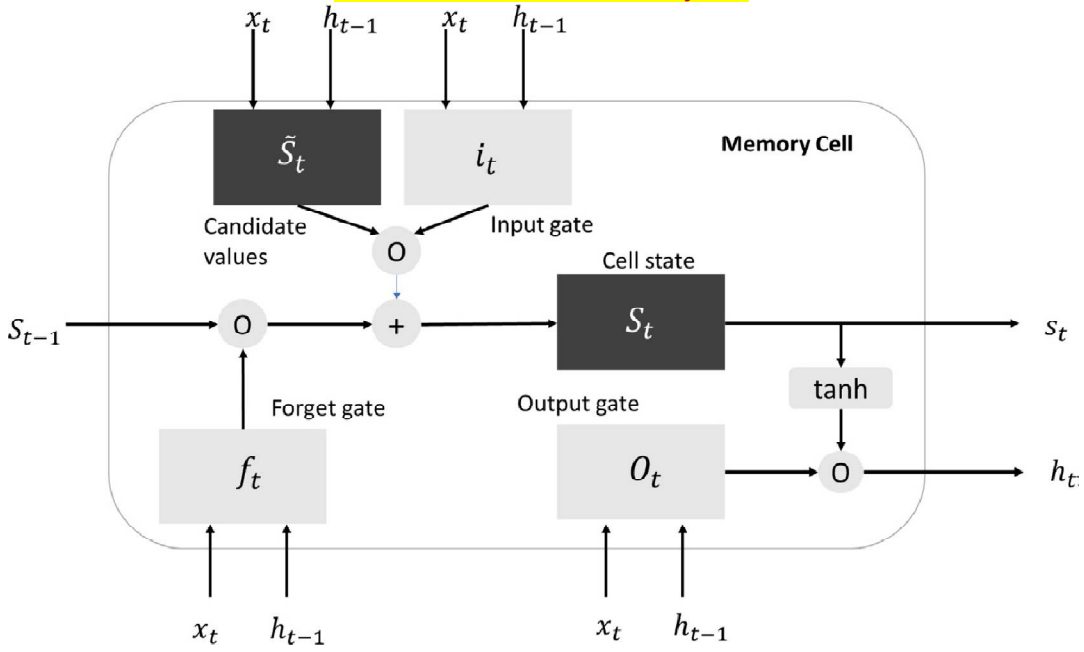


Generating feature image masks from last layer activation's of CNN model F

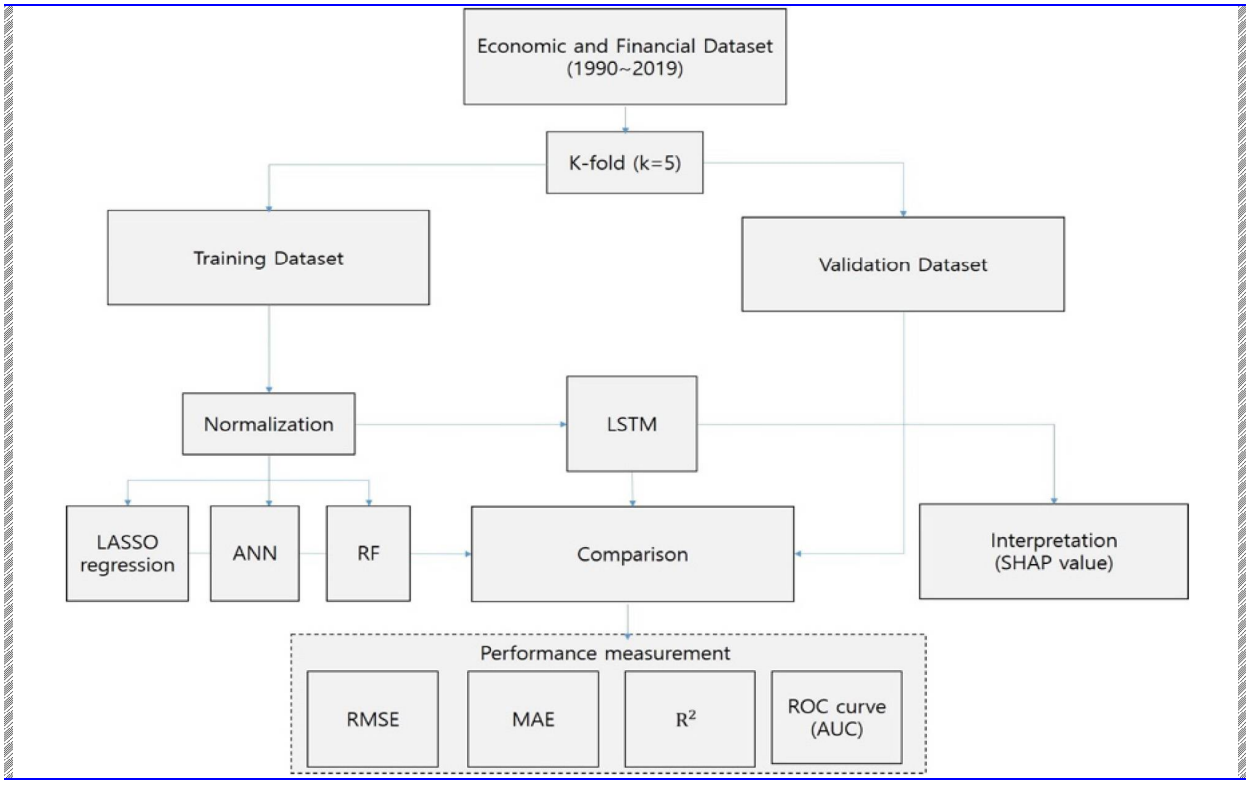


CNN model F is same of all the steps

Structure of an LSTM memory cell



Flow diagram of prediction process in models



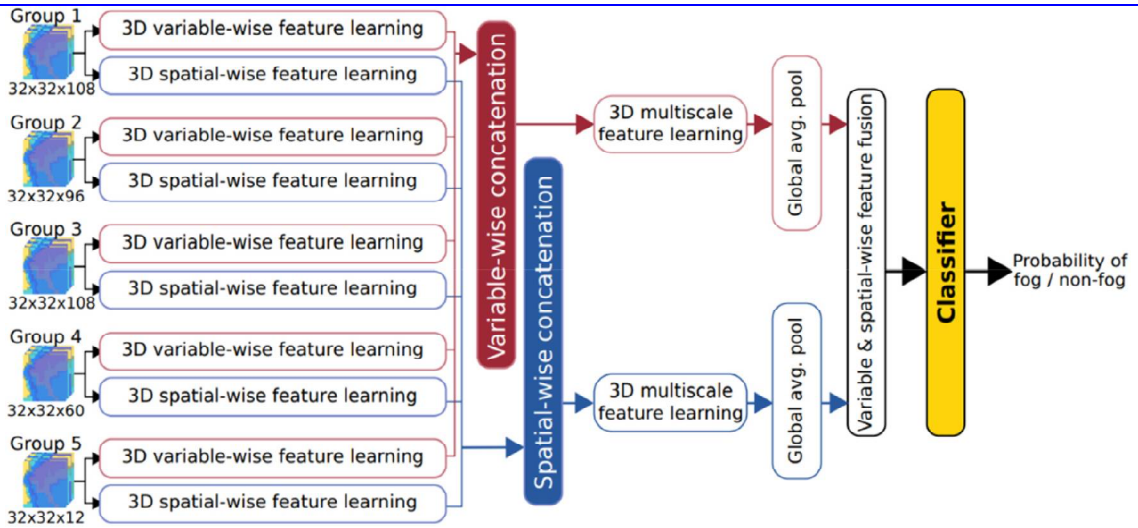
xAI.		2022-127

Principal hyperparameters of Denoising Generative Adversarial Network.

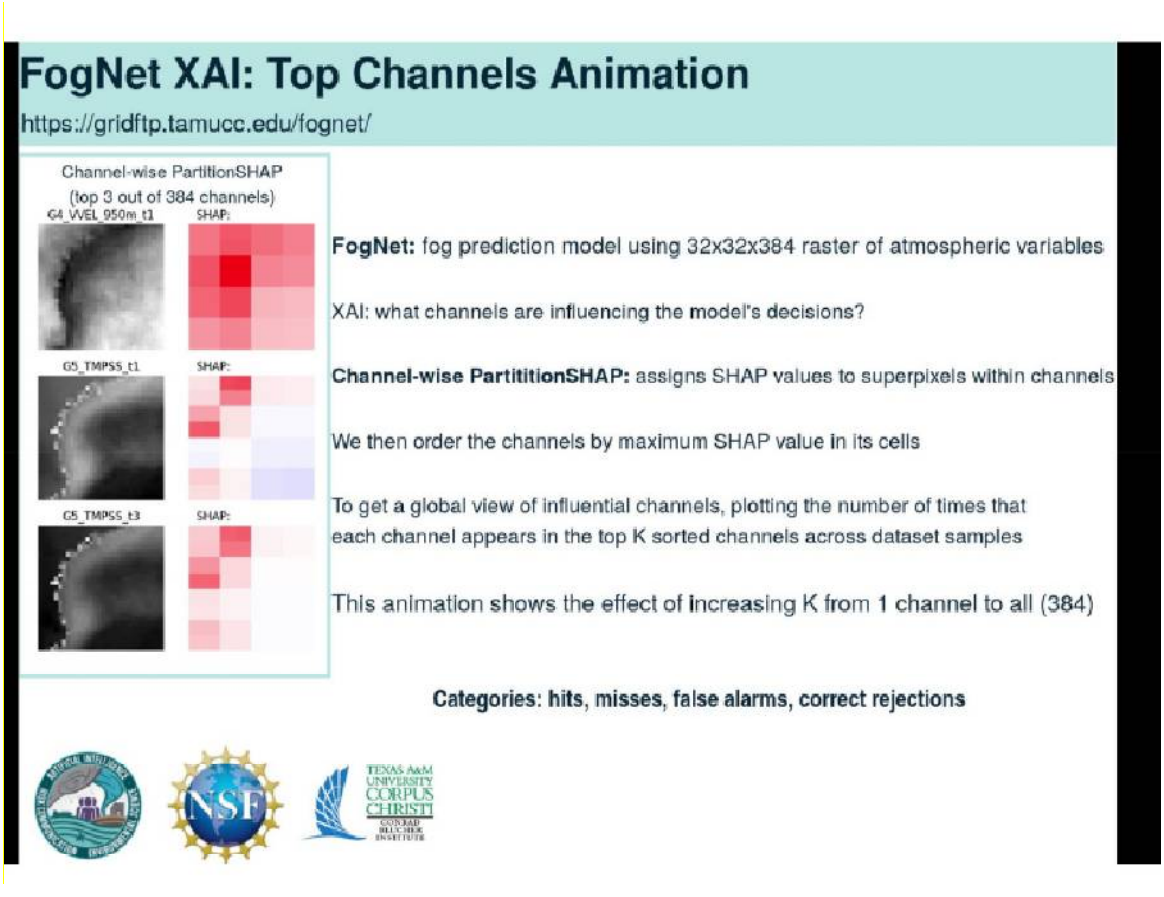
Parameter	Value
2D representation width	20
2D representation height	20
2D number of channels	1
Base number of filters	64
Kernel size	(3, 3)
Strides	(2, 2)
Dropout rate	0.5
Learning rate	2e-4
Max epochs	2000
Regularization L1 lambda	100

xAI.		2022-138

Overview of the FogNet3D parallel processing of features – spatial-wise (blue) and variable-wise (red)

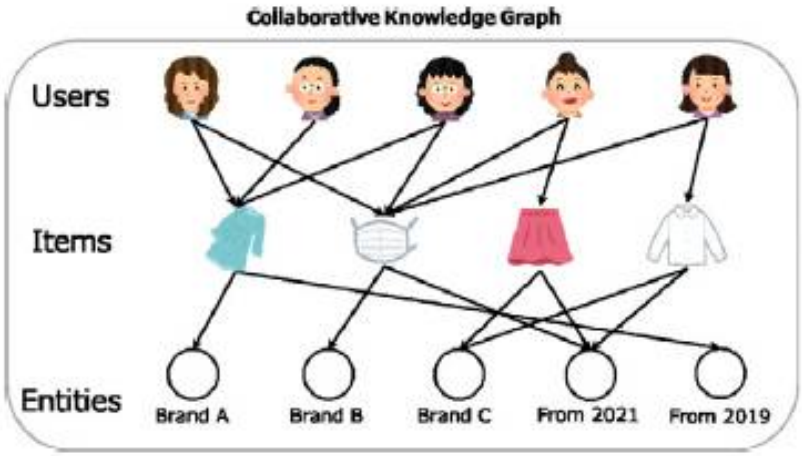


Animation showing the occurrences of each channel in the top-K channels
[K :1, 2, …, 384]



xAI.	Collaborative Knowledge Graph (CKG)	2022-141

CKG embedding layer



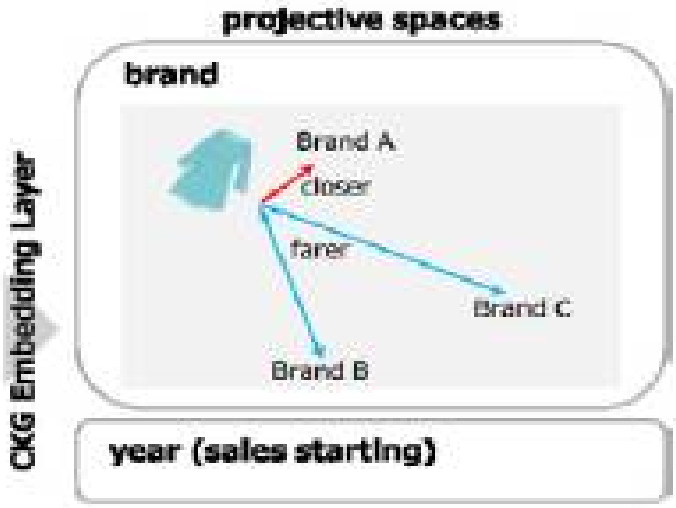


Illustration of attentive embedding propagation layer

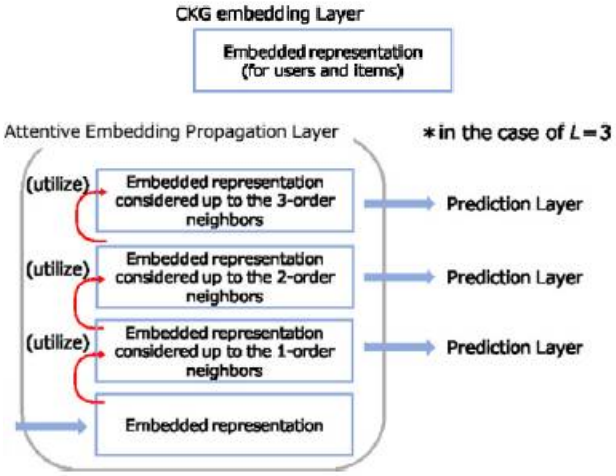


Illustration of prediction layer

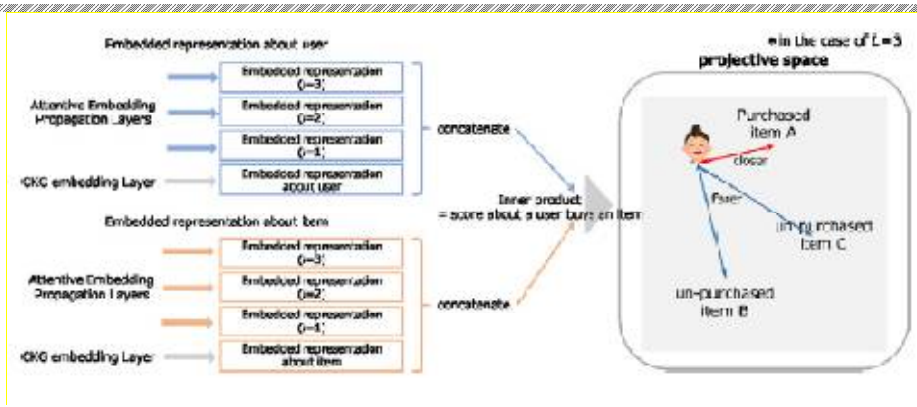
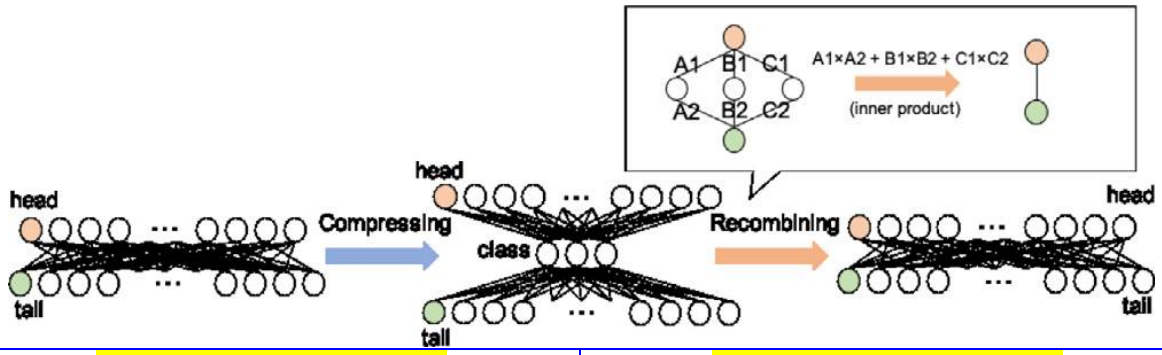
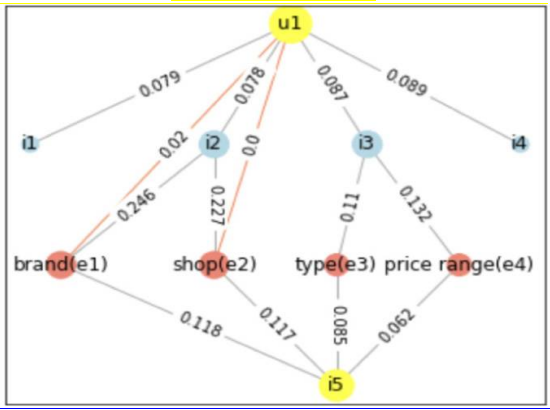


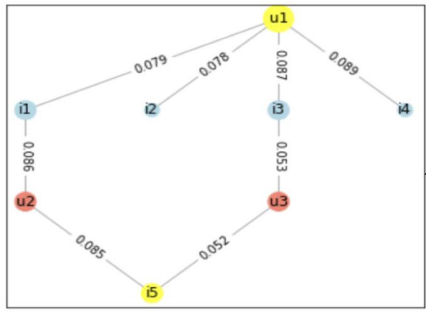
Illustration of compressing and recombining

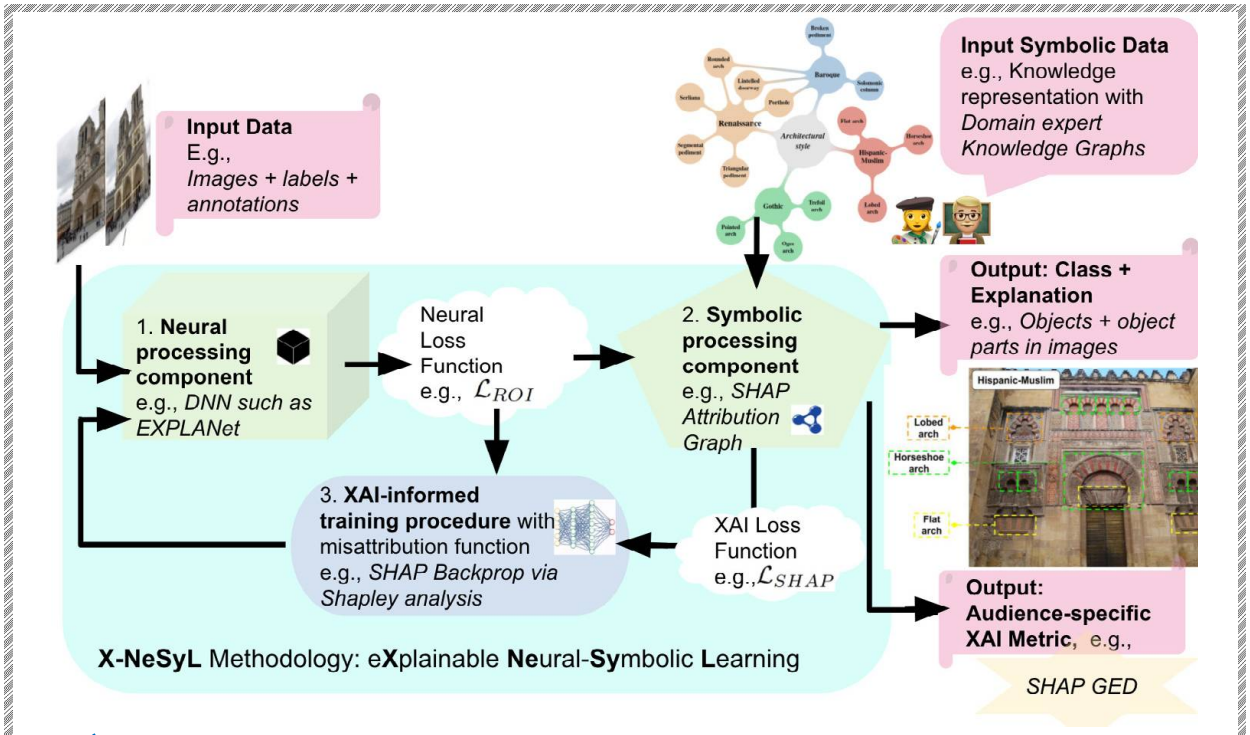


Example of a map visualizing attribute-based reason for recommendation.



Example of a map visualizing behavior-based reason for recommendation.



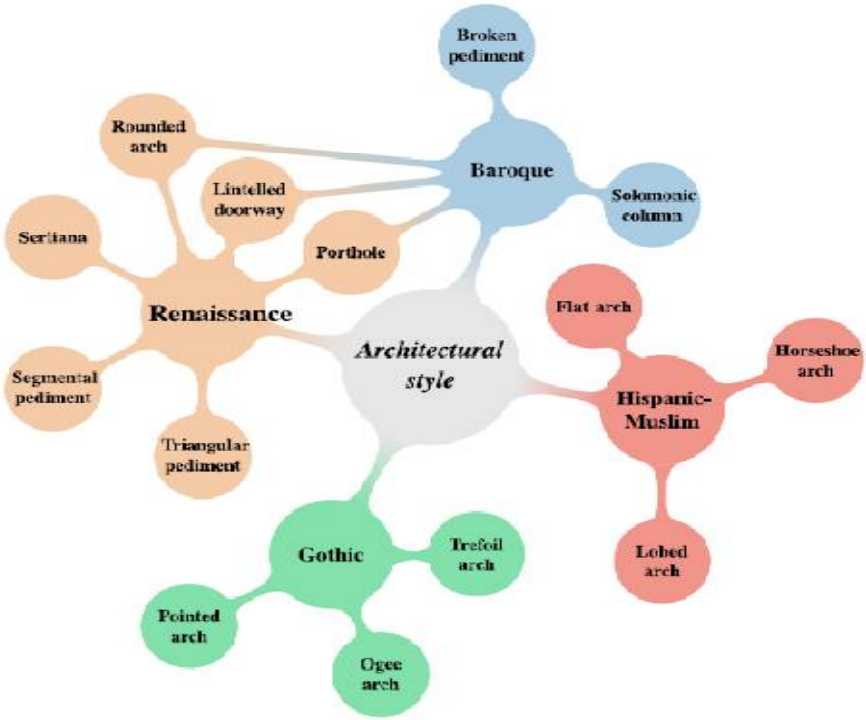


- ✓ Object detection and classification model
- ✓ Knowledge graphs
- ✓ SHAP-Backprop training procedure

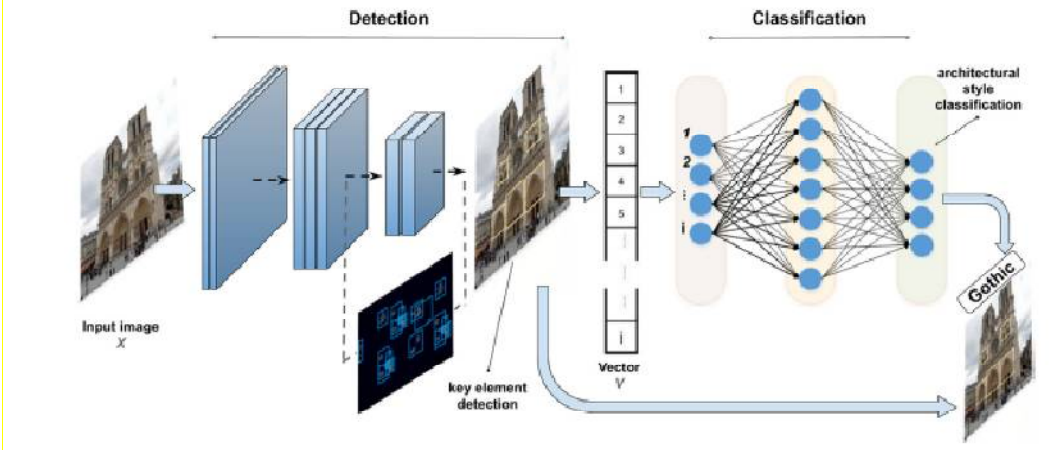
xAI.		2022-

Simplified Monu-MAI

knowledge graph constructed based on art historian's expert knowledge

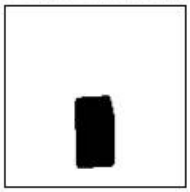
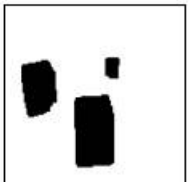


EXPLANet architecture processing examples from MonuMAI dataset



Example CLEVR-XAI-complex data point

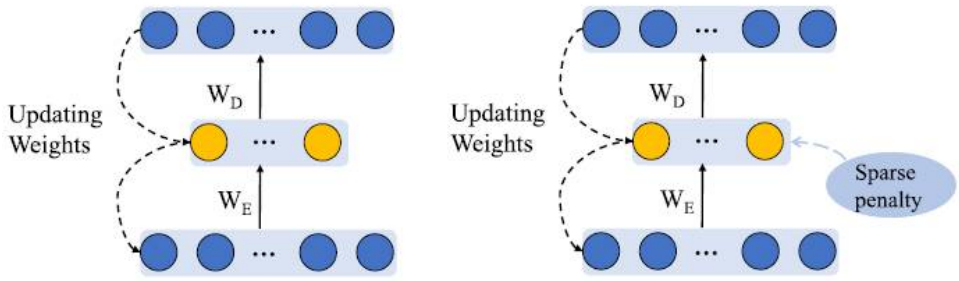
Image	Question/Answer	Program
	What number of objects are large purple blocks or green metallic cubes? *one*	branch1 = [scene, filter_size, filter_color, filter_shape] branch2 = [scene, filter_color, filter_material, filter_shape] program = [(branch1,branch2), union, count]

Ground Truth	GT Unique	GT Unique First-non-empty
	undefined	
	GT Union	GT All Objects
		

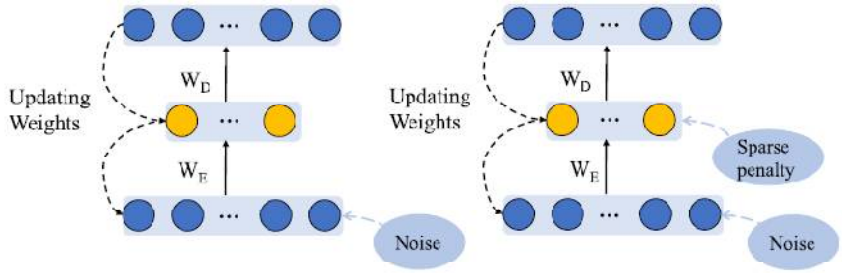
Objective: Which objects in the scene are considered as ground truths

xAI.		2022-154

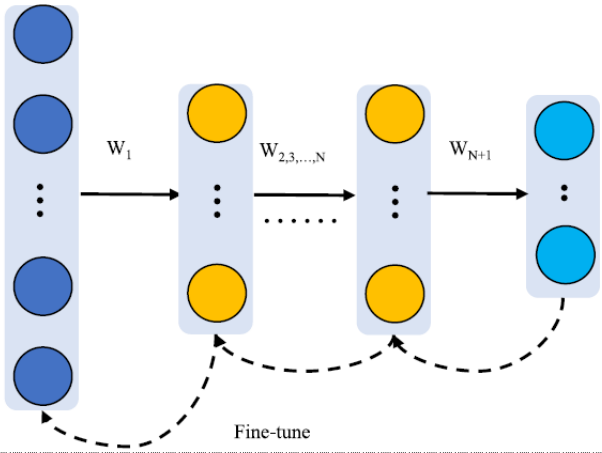
Structure of autoencoder, sparse autoencoder,
denoising autoencoder, denoising sparse autoencoder (DSAE)



Autoencoder Sparse autoencoder



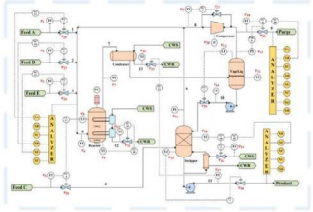
Supervised fine-tuning process of DSAE

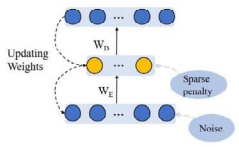


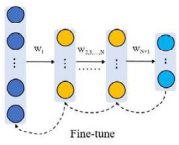
xAI.		2022-154

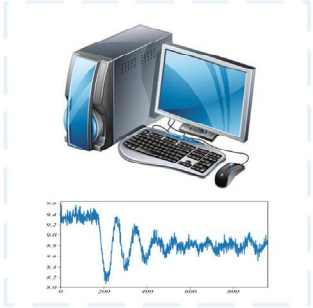
Framework of robust and understandable fault detection and diagnosis

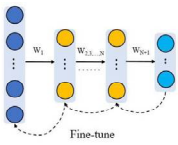
DSAE-SIG (smooth integrated gradients)

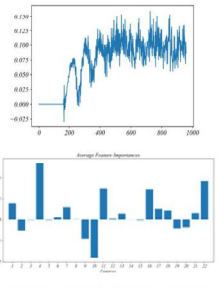






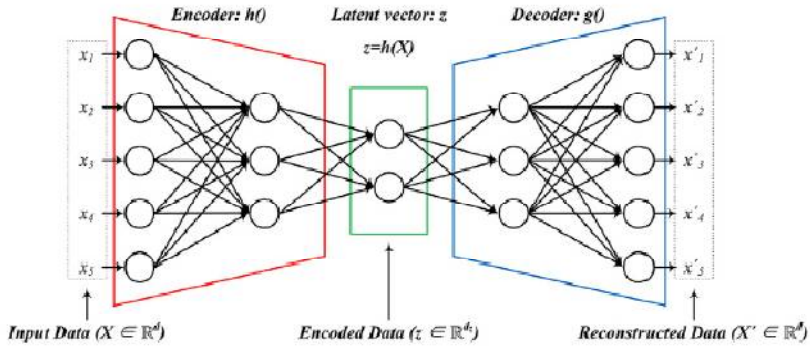




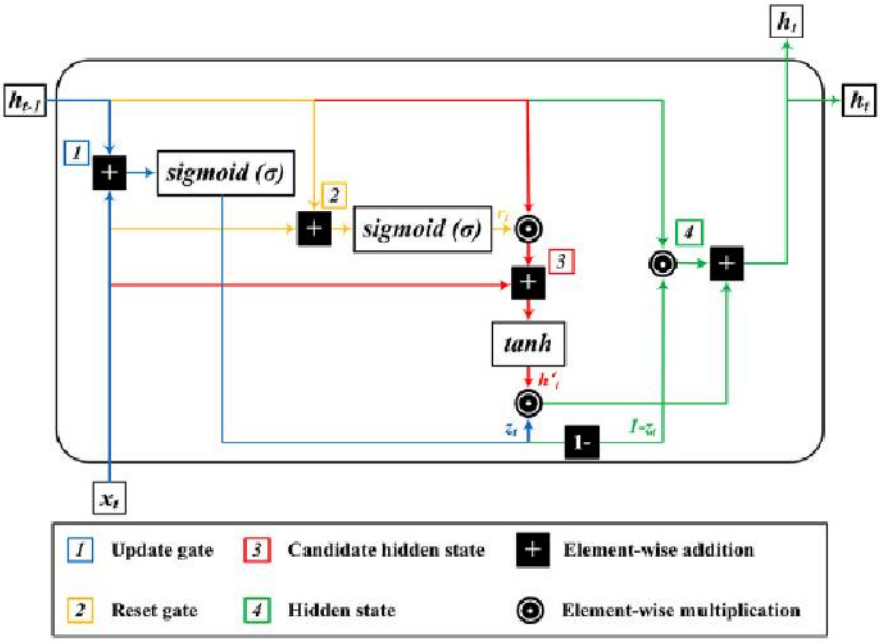


Scenario	Method name
Binary Fault Detection and Diagnosis	PCA KPCA Two-class SVM AE SAE DAE RNPCA DSAE
Multi-class Fault Detection and Diagnosis	SVM AE SAE DAE DSAE

Structure of AE method



Structure of GRU network



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$$

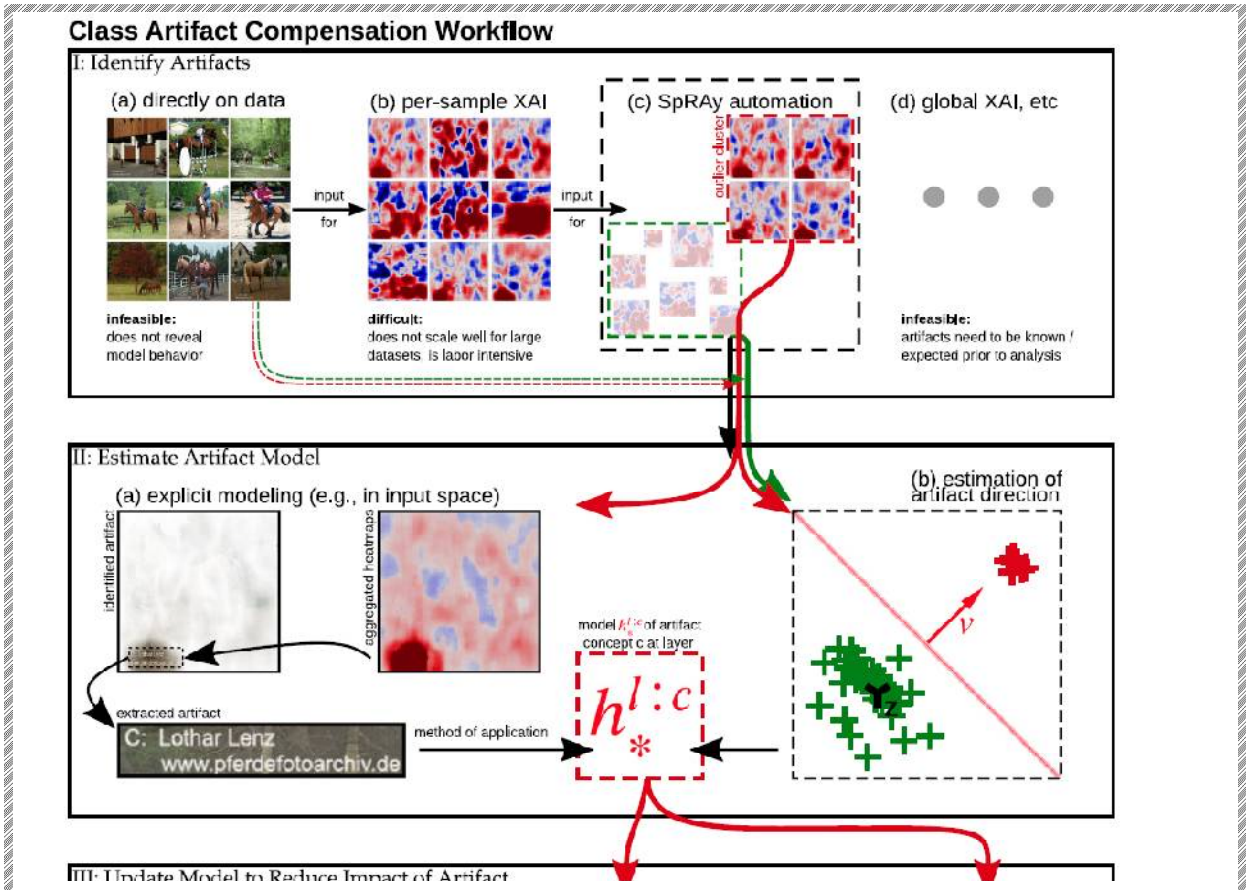
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$

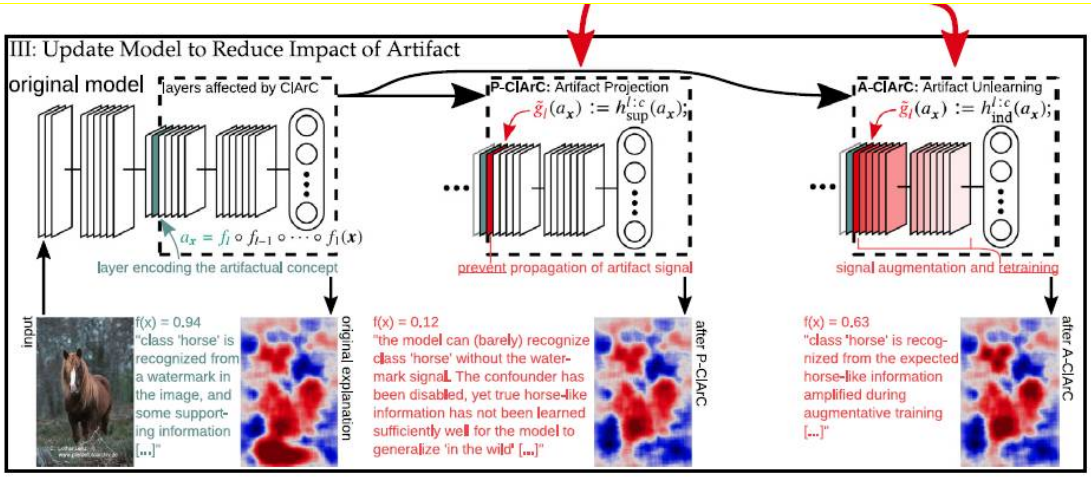
$$\hat{h}_t = tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h)$$

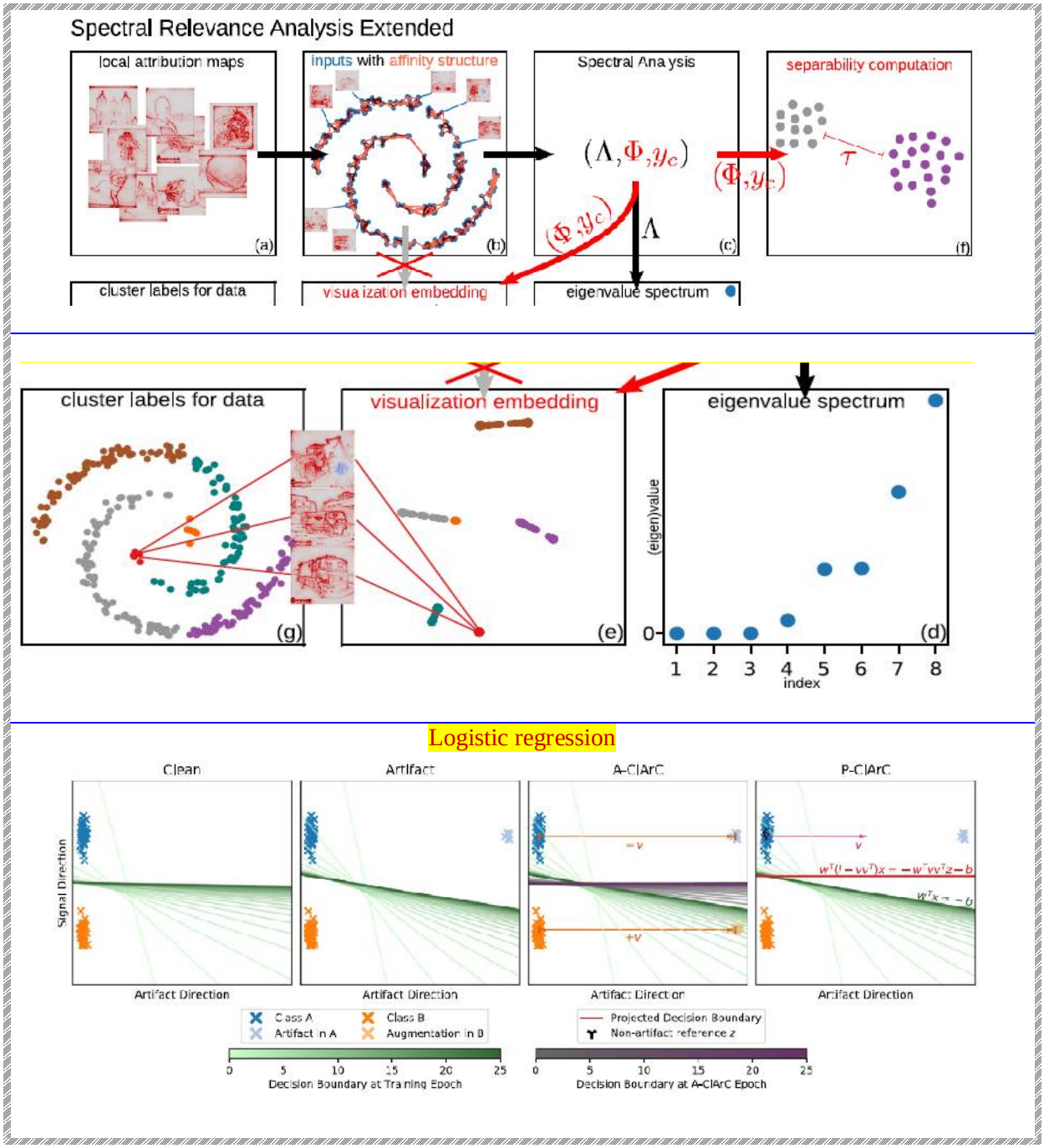
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t$$

xAI.		2022-157

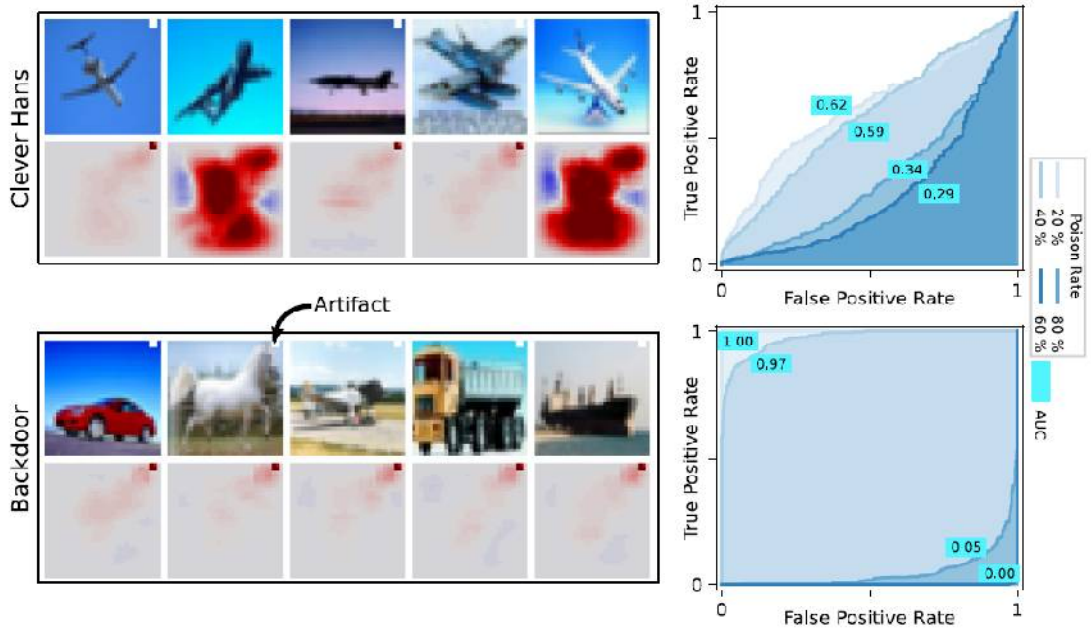
Class Artifact Compensation framework

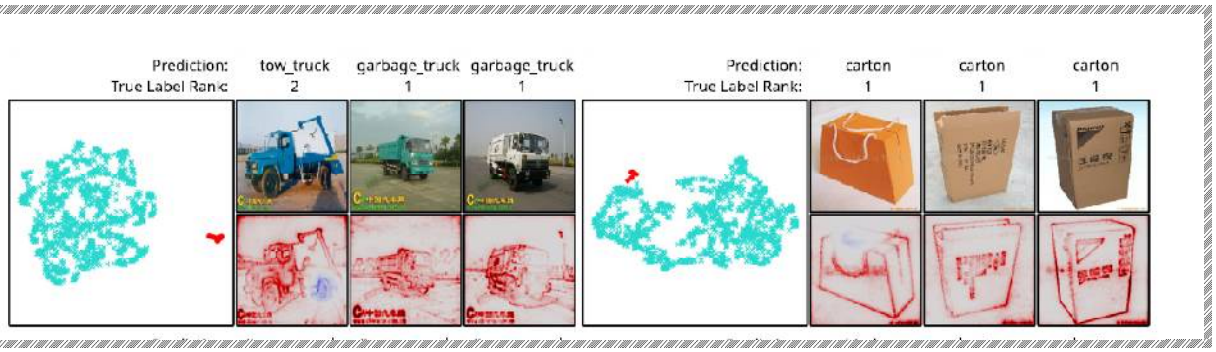

Class Artifact Compensation Workflow
I: Identify Artifacts
(a) directly on data
(b) per-sample XAI
(c) SpRAy automation
(d) global XAI, etc
input for
input for
outlier cluster
infeasible:
does not reveal model behavior
difficult:
does not scale well for large datasets & is labor intensive
infeasible:
artifacts need to be known / expected prior to analysis
II: Estimate Artifact Model
(a) explicit modeling (e.g., in input space)
(b) estimation of artifact direction
identified artifact
aggregated heatmaps
model $h^{l:c}_*$ of artifact concept c at layer
extracted artifact
C: Lothar Lenz
www.pferdefotoarchiv.de
method of application
$h^{l:c}_*$
v
III: Update Model to Reduce Impact of Artifact

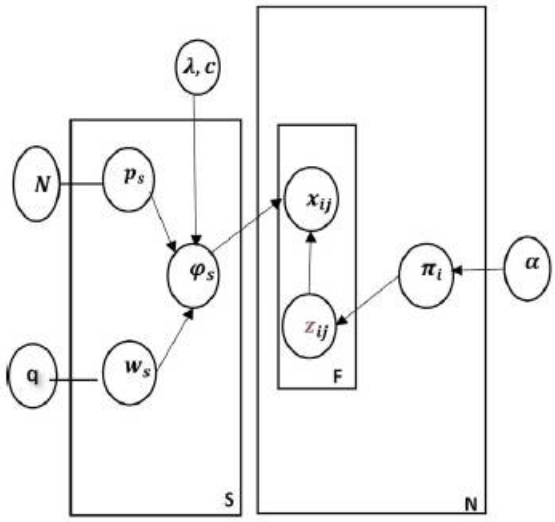

III: Update Model to Reduce Impact of Artifact
original model
layers affected by ClArC
P-ClArC: Artifact Projection
A-ClArC: Artifact Unlearning
$a_x = f_l \circ f_{l-1} \circ \cdots \circ f_1(x)$
layer encoding the artifactual concept
prevent propagation of artifact signal
signal augmentation and retraining
input
f(x) = 0.94
"class 'horse' is recognized from a watermark in the image, and some supporting information [...]"
original explanation
f(x) = 0.12
"the model can (barely) recognize class 'horse' without the watermark signal. The confounder has been disabled, yet true horse-like information has not been learned sufficiently well for the model to generalize 'in the wild' [...]"
after P-ClArC
f(x) = 0.63
"class 'horse' is recognized from the expected horse-like information amplified during augmentative training [...]"
after A-ClArC



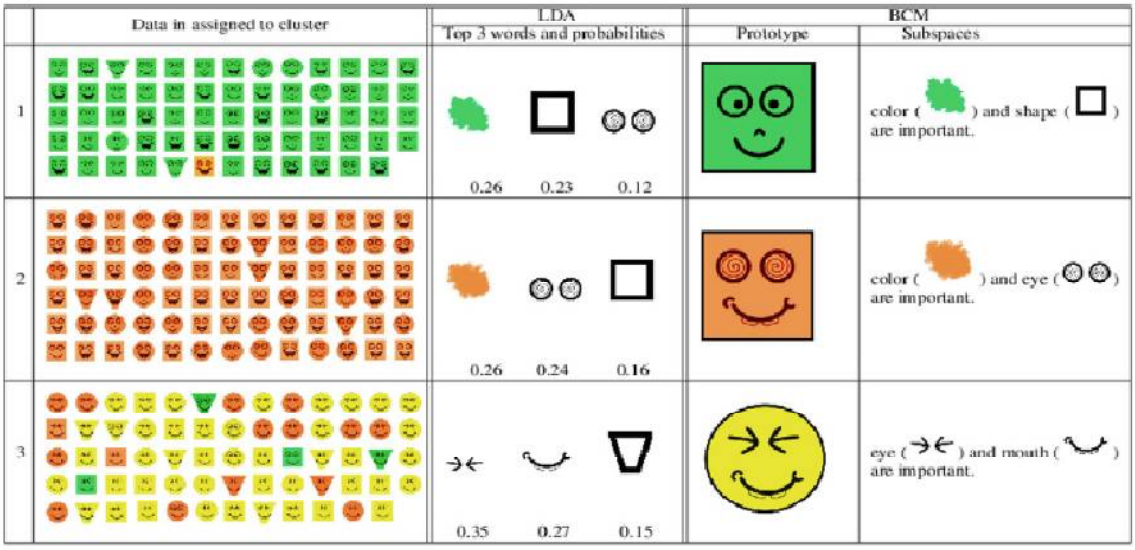
Differences in detection of
CH artifacts (top) and BDs (bottom)







(a) Flow of Bayesian Case Model (BCM)



(b) BCM and LDA interpretation for dataset of smiley faces [77]