| | **Paper** | | | | |
|---|---|---|---|---|---|
| ***Ontology*** | | | | | |
| explanationExperience.rdf | **ExplanationExperience** | | | | |
| | hasDescription | **Description** | | | |
| aimodel.rdf | | hasAIModel | **AIModel** | | |
| | | | trained on | **Dataset** | |
| | | | | hasDataType | **DataType** |
| | | | | | **number of features** |
| | | | | | **number of instances** |
| | | | solves | **AITask** | |
| | | | | hasType | **AITaskType** |
| | | | | hasGoal | **AITaskGoal = Description** |
| | | | utilises | **AIMethod** | |
| | | | | hasType | **AIMethodType** |
| aimodelevaluation.rdf | | | annotated by | AIModelAssessmentResult | |
| | | | | basedOn | **AIModelAssessmentMetric** |
| | | | | measures | **AIModelAssessmentDimension** |
| explainer.rdf | | hasExplainer (needExplainer?) | **Explainer** | | |
| | | | | *hasOutputType* | ***Explanation*** |
| | | | | hasPortability | **Portability** |
| | | | | hasConcurrentness | **ExplainerConcurrentness** |
| | | | | hasPresentation | **InformationContentEntity** |
| | | | | *hasExplanationScope* | ***Explanation Scope*** |
| | | | | targetType | **Explanation Target** |
| user.rdf | | hasUser | **User** | | |
| | | | asks | **UserQuestion** | |
| | | | | hasTarget | **UserQuestionTarget** |
| | | | | hasType | **QuestionType** |
| | | | has intent | **Intent** | |
| | | | | hasType | IntentType |
| | | | has resources | **Technical Facilities** | |
| | | | | can handle | **ExplanationModality** |
| | | | hasGoal | **process** | |
| | | | possess | **Domain Knowledge** | |
| | | | | level of domain knowledge | **KnowledgeLevel** |
| | | | | level of AI knowledge | **KnowledgeLevel** |
| behaviour_tree.rdf | hasSolution | **Solution** | | | |
| | | hasExplainer | **Explainer** | | |
| | | | utilises | **ExplainabilityTechnique** | |
| | | | | hasType | **ExplainabilityTechniqueType** |
| | | | | hasOutputType | **Explanation** |
| | | | | hasPortability | **Portability** |
| | | | | hasConcurrentness | **ExplainerConcurrentness** |
| | | | | hasPresentation | **InformationContentEntity** |
| | | | | *hasExplanationScope* | ***Explanation Scope*** |
| | | | | targetType | **Explanation Target** |
| | | | | isCompatiblewithFeatureTypes | **DataType** |
| | | | | hasComplexity | **ComputationalComplexity** |
| userevaluation.rdf | hasDone | **UserEvaluation** | | | |
| | | basedOn | **Metric** | | |
| | | | measures | **Dimensions** | |
| **Code Availability** | | | | | |

| | |
|---|---|
| Make it personal: a social explanation system applied to group recommendations | *DisCERN: Discovering Counterfactual Explanations using Relevance Features from Neighbourhoods* |
| ***PSIE*** | ***DisCERN*** |
| PsieGroupRecommendationExplanationExperience | LungCancerRiskExplanationExperience |
| Recommend movies to groups based on social knowledge | predict lung cancer risk given clinical data of patients |
| HappyMovieRecommenderSystem | LungCancerRiskPredictionModel |
| HappyMovieDataset | LungCancerRiskDataset |
| Tabular | Tabular |
| N/A | 12 |
| N/A | 427 |
| GroupRecommendation | CancerRiskPrediction |
| Recommendation | Multi-class Classification |
| Recommend movies to groups based on social knowledge | predict lung cancer risk given clinical data of patients |
| SocialGroupRecommender | CancerRiskRandomForest |
| Knowledge based Recommender | RandomForest |
| 3.86, 3.69, 3.56, 3.89 | 87.94 |
| Likert scale (5 points - being 5 strongly agree) | Accuracy |
| Usefulness, Decision process, Reusability, Usability | Performance |
| | |
| Content Based Explanation | Counterfactual Explanation |
| model-specific | model-agnostic |
| post-hoc | post-hoc |
| Any | Computational Entity |
| local | local |
| prediction | prediction |
| Moviegoer | Patient/Clinician |
| Why does the system recommend movie Y for group X? | How can patient X reduce cancer risk Y to lower |
| System Recommendation | System Recommendation |
| Why Question | How/What-if Question |
| Understand system recommendation | Reducing cancer risk |
| Trust, Satisfaction | Education, Taking Action |
| ScreenDisplay | ScreenDisplay |
| Any | Any |
| | |
| User profile | Any/ Clinical knowledge |
| high | low/high |
| low | low |
| | |
| | DisCERNCancerRiskExplainer |
| PsieRuleBasedTechnique | DisCERN |
| Knowledge Extraction | Feature Relevance+Example based |
| ContentBasedExplanation | Counterfactual Explanation |
| model-specific | model-agnostic |
| post-hoc | post-hoc |
| Visual/Textual | Computational Entity |
| local | local |
| prediction | prediction |
| Tabular | Tabular |
| N/A | N/A |
| | N/A |
| Questionnaire | N/A |
| Usefulness/Helpfulness | N/A |
| N/A | https://github.com/RGU-Computing/DisCERN-XAI |

| | |
|---|---|
| Evaluating Explainability Methods Intended for Multiple Stakeholders | Directing exploratory search: Reinforcement learning from user interactions with keywords |
| ***BTTelecom*** | ***SciNet*** |
| BTTelecomRecommenderExplanationExperience | DocumentSearchExplanationExperience |
| recommend engineering notes to desk support staff to help on-site engineers | Determine the most related documents given a set of keywords |
| EngineerNoteRecommender | SciNetSearchEngine |
| BTEngineeringNotes | WebOfScienceDataset |
| Text | Documents |
| 300 tf-idf features | 7 things, title, abstract, author names, publication year, publication forum, article, keywords |
| 5352 | 50million |
| EngineerNoteRecommendation | DocumentRetrieval |
| Recommendation,classification | InformationRetrieval |
| Predict next scenario based on description in the engineering notes | Determine the most related documents given a set of keywords |
| EngineerNoteRecommender | SciNetReinforcementLearning |
| Content based Recommender, Machine Learning / term frequency, unsupervised | Reinforcement Learning |
| 50.88%, 99.10%(in lab), completeness - 70% (in practice) | 0.71 |
| Accuracy with and without (NNR) class, top K Accuracy, confidence score | Kappa |
| Performance, efficiency of scenario organisation | agreement between expert and system |
| | |
| Neighbourhood Explanation, numerical, textual | Neighbourhood Explanation |
| model-agnostic | model-specific |
| post-hoc | ante-hoc |
| Any | visual |
| local | local |
| prediction | prediction |
| Desk Agent | Scientist |
| Why task Y is recommended as next task? | Why was this result X retrieved for this query Y? |
| System Recommendation | System Recommendation |
| Why question | Why Question |
| why a recommendation has been made? | Understand system prediction |
| Transparency, Taking Action, Education | Effectiveness, Satisfaction |
| ScreenDisplay | ScreenDisplay |
| Text, Image | Any |
| | |
| BTNetworkPlannerDomainExpert, BTFieldEngineerDomainExpert, BTDeskAgentDomainNovice | Search Domain |
| high, high, low | high |
| low | low |
| | |
| BTRecommenderExplainer | |
| BTContentSimilarityBasedTechnique: confidence score, feature-importance, summarisation of sim/difs | SciNetReinforcementLearning |
| Knowledge Extraction + Feature Relevance | Knowledge Extraction |
| Neighbourhood Explanation | Neighbourhood Explanation |
| model-agnostic | model-specific |
| post-hoc | ante-hoc |
| Content + Similarity | Interactive visual |
| local | local |
| prediction | prediction |
| Text | Documents |
| N/A | N/A |
| see Notes | |
| Question to get feedback | Questionnaire |
| Usefulness/Educatingness/Efficiency | Usability/Quality of user experience |
| N/A | N/A |

| | |
|---|---|
| Visualizing Recommendations to Support Exploration, Transparency and Controllability | Axiomatic Attribution for Deep Networks |
| ***TalkExplorer*** | ***IntGradImage*** |
| TalkExplorerExplanationExperience | IGImageClassificationExplanationExperience |
| Recommend papers based on content and social connections | predict the category of a given image |
| ConferenceNavigator3RecommenderSystem | IGImageClassificationModel |
| ConferenceNavigator3Dataset | ILSVRC-2014 |
| Tabular | Image |
| N/A | 89401 pixels |
| N/A | 456182 |
| TalkPaperRecommendation | ImageClassification |
| Recommendation | Multi-class Classification |
| Recommend papers based on content and social connections | predict the category of a given image |
| CN3ContentBasedRecommender | GoogleNet |
| Tf-idf + kNN | Convolutional Neural Network |
| N/A | 6.67% |
| N/A | top-5 error |
| N/A | Performance |
| | |
| Neighbourhood Explanation | Saliency Map Explanation |
| model-specific | model-specific |
| ante-hoc | post-hoc |
| Image | Image |
| local | local |
| prediction | prediction |
| Conference atendee | Any User |
| Why does the system recommend paper Y to user X? | Why does the system predict category Y for image X? |
| System Recommendation | System Recommendation |
| Why Question | Why Question |
| Understand system prediction | Understand how system works |
| Effectiveness, Transparency, Scruitability | Transparancy, Trust |
| ScreenDisplay | ScreenDisplay |
| Any | Any |
| | |
| Conference Topic Domain | Public Domain |
| high | Any |
| low | Any |
| | |
| | |
| TalkExplorerKNNTechnique | IntegratedGradientTechnique |
| k-nearest Neighbour | IntegratedGradient |
| Neighbourhood Explanation | Saliency Map Explanation |
| model-specific | model-specific |
| ante-hoc | post-hoc |
| visual | Annotated Computational Entity |
| local | local |
| prediction | prediction |
| Tabular | Image |
| N/A | N/A |
| Questions about explanation visualisation knowledge + tasks with TalkExplorer + Likert Scale questions about their needs | N/A |
| Think Aloud + Likert Scale | N/A |
| Effectiveness | N/A |
| N/A | https://github.com/ankurtaly/Integrated-Gradients |

| | |
|---|---|
| Axiomatic Attribution for Deep Networks | Axiomatic Attribution for Deep Networks |
| ***IntGradRetinopathy*** | ***IntGradTextClassification*** |
| DiabeticRetinopathyDetectionExplanationExperience | QuestionCategoryExplanationExperience |
| predict if a given medical image contains diabetic retinopathy | predict the question category based on question text |
| DiabeticRetinopathyDetectionModel | KimQuestionCategoryPredictionModel |
| EyePACS | WikiTableQuestions dataset |
| Image | text |
| 1382400 pixels | N/A |
| 128175 | 22033 |
| DiabeticRetinopathyDetection | QuestionCategoryPrediction |
| Binary Classification | Multi-class Classification |
| predict if a given medical image contains diabetic retinopathy | predict the question category based on question text |
| Fine-tubed InceptionV3 | KimCNN-multichannel |
| Convolutional Neural Network | Convolutional Neural Network |
| 90.3%, 98.1%, 99.1% | N/A |
| FOCP Sensitivity, FOCP Specificity, AUROC for EyePACS | N/A |
| Performance | N/A |
| | |
| Saliency Map Explanation | Saliency Map Explanation |
| model-specific | model-specific |
| post-hoc | post-hoc |
| Image | Text |
| local | local |
| prediction | prediction |
| Clinician, Optomologist | Any User |
| Why does the system predict RDR for image X? | Why does the system predict category Y for question text X? |
| System Recommendation | System Recommendation |
| Why Question | Why Question |
| Understand how system works | Understand how system works |
| Transparancy, Trust, Education | Transparancy, Trust |
| ScreenDisplay | ScreenDisplay |
| Any | Any |
| | |
| Clinical Knowledge | Public Domain |
| High | low |
| Low | low |
| | |
| | |
| IntegratedGradientTechnique | IntegratedGradientTechnique |
| IntegratedGradient | IntegratedGradient |
| Saliency Map Explanation | Saliency Map Explanation |
| model-specific | model-specific |
| post-hoc | post-hoc |
| Annotated Computational Entity | Annotated Computational Entity |
| local | local |
| prediction | prediction |
| Image | Text |
| N/A | N/A |
| N/A | N/A |
| N/A | N/A |
| N/A | N/A |
| https://github.com/ankurtaly/Integrated-Gradients | https://github.com/ankurtaly/Integrated-Gradients |

| | |
|---|---|
| Textual Explanations for Self-Driving Vehicles | iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction |
| ***KimEtAlMethod*** | ***iBCM*** |
| SelfDrivingExplanationExperience | iBCMGradingExplanationExperience |
| make acceleration or change course decisions in a self-driving car based on video | Cluster student assignment submissions to design grading rubric or to compose feedback |
| Self-drivingDecisionMakingModel | iBCMClusteringModel |
| BerkeleyDeepDriveDataset | iBCMAssessmentsDataset |
| Video | Code |
| 40 seconds (frame rate not known) | N/A |
| 6984 | N/A |
| Self-drivingVehicalControl (acceleration and course) | AssessmentsClustering |
| Regression | Clustering |
| make acceleration or change course decisions in a self-driving car based on video | |
| Deep Neural Networks with Attention | iBCMClusteringMethod |
| NeuralNetwork | BCM |
| [2.29, 0.82], [6.06, 0.47] | N/A |
| [Mean of absolute error, Mean of distance correlation] of Acceleration and Course | N/A |
| self-drivingVehicalControl performance | N/A |
| | |
| Introspective Explanation | Prototype Explanation |
| model-agnostic | model-specific |
| post-hoc | ante-hoc |
| text | Computational Entity |
| local | cohort |
| prediction | prediction |
| Driver | Lecturer |
| Why does the vehical system make decision X? | Why does the system assign certian assessments in to one cluster? How does the system assign clusters? |
| System Recommendation | Model |
| Why Question | Why/How Question |
| Understand system decision | Understand system/Understand cohort of predictions |
| User acceptance, Trust, Understanding, Effective communication | Education/Transparency |
| ScreenDisplay | ScreenDisplay |
| text | Any |
| | |
| Public Domain | Lecturer Knowledge |
| high | High |
| low | Low |
| | |
| LSTMTextGeneratorExplanationTechnique | iBCMTechnique |
| Data-driven Explanation Generation | Interactive Bayesian Case Model |
| Introspective Explanation | Prototype Explanation ,Feature Importance Explanation |
| model-agnostic | model-specific |
| post-hoc | ante-hoc |
| text | Visual, textual |
| local | cohort |
| prediction | prediction |
| Any | Tabular,Image,Text |
| N/A | N/A |
| N/A | N/A |
| N/A | Questionnaire |
| N/A | Usefulness/Efficiency |
| https://github.com/pair-code/saliency | N/A |

A case-based reasoning system for aiding detection and classification of nosocomial infections

***InNoCBR***

InnoInfectionExplanationExperience
Predict patient's infection based on a clinical, laboratory, and medico administrative based data
InnoInfectionDiagnosisModel
InnHostpitalDataset
Tabular
6 tuple {S, B, V, S, C, E}
5385
InfectionDiagnosis
Multi-class Classification
Predict patient's infection based on a clinical, laboratory, and medico administrative based data
InnoHybridMethod
Rules+PARTRules+NLP(NB)
70.21%, 0.62, 55.75%, 19.18%
Accuracy, Kappa, false-positive rate (before and after modified data)
performance

Reasoning Path Explanation
model-specific
ante-hoc
text
local
prediction
Spanish NHS Doctor
Why does the system predict infection Y for patient X?
System Recommendation
Why Question
Validate system prediction
Trust/Transparency
ScreenDisplay
Any

Infection Detection/Prevention Knowledge
high
low

InnoDecisionPathTechnique
Decision tree
Reasoning Path Explanation
model-specific
anti-hoc
Text
local
prediction
Tabular
N/A
N/A
N/A
N/A
N/A

Explaining Models by Propagating Shapley Values

***DeepSHAPGlobal***

MortalityPredictionExplanationExperience
Predict patient mortality base on clinical, nutritional and behaviorial factors
MLPMortalityPredictionModel
NHANESDataset
Tabular
79
14407
MortalityPrediction
Binary Classification
Predict patient mortality base on clinical, nutritional and behaviorial factors
MoralityMLP
Neural Network
82.56%
Accuracy
Performance

Feature Importance Explanation
model-agnostic
post-hoc
Any
global
model
Clinician
What/How features contributed to predicting mortality Y for patient X?
Model
What/How Question
Understand how model make decisions
Transparency
ScreenDisplay
Any

Clinical Knowledge
high
low

DeepSHAPExplanationTechnique
SHAP/Game-theory
Feature Importance Explanation
model-agnostic
post-hoc
violin plot chart
global
model
Tabular
N/A
N/A
N/A
N/A
https://github.com/lrjball/shap

Explaining Models by Propagating Shapley Values

***DeepSHAPLocal***

MortalityExplanationExperience

Predict patient mortality base on clinical, nutritional and behaviorial factors

MLPMortalityPredictionModel

NHANESDataset

Tabular

79

14407

MortalityPrediction

Binary Classification

Predict patient mortality base on clinical, nutritional and behaviorial factors

MoralityMLP

Neural Network

82.56%

Accuracy

Performance

Feature Importance Explanation

model-agnostic

post-hoc

Any

local

prediction

Clinician

What/How features contributed to predicting mortality Y for patient X?

System Recommendation

What/How Question

Understand why model made a decision

Transparency/Education

ScreenDisplay

Any

Clinical Knowledge

high

low

DeepSHAPExplanationTechnique

SHAP/Game-theory

Feature Importance Explanation

model-agnostic

post-hoc

bar chart

local

prediction

Tabular

N/A

N/A

N/A

N/A

https://github.com/lrjball/shap