UNIVERSITÀ DI PISA

XAI Models for Quality of Experience Prediction in Wireless Networks

Alessandro Renda, Pietro Ducange, Gionatan Gallo, Francesco Marcelloni

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CROSSLAB Innovation for industry 4.0

Hexa-X

FUZZ-IEEE 2021 LUXEMBOURG

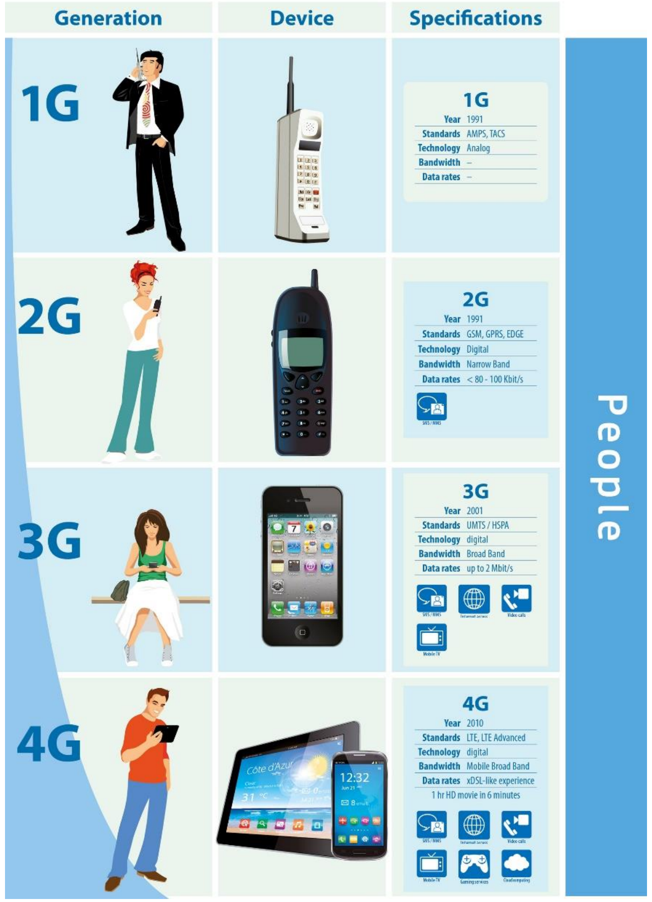
HANDLING UNCERTAINTY IN INTERPRETABLE AI

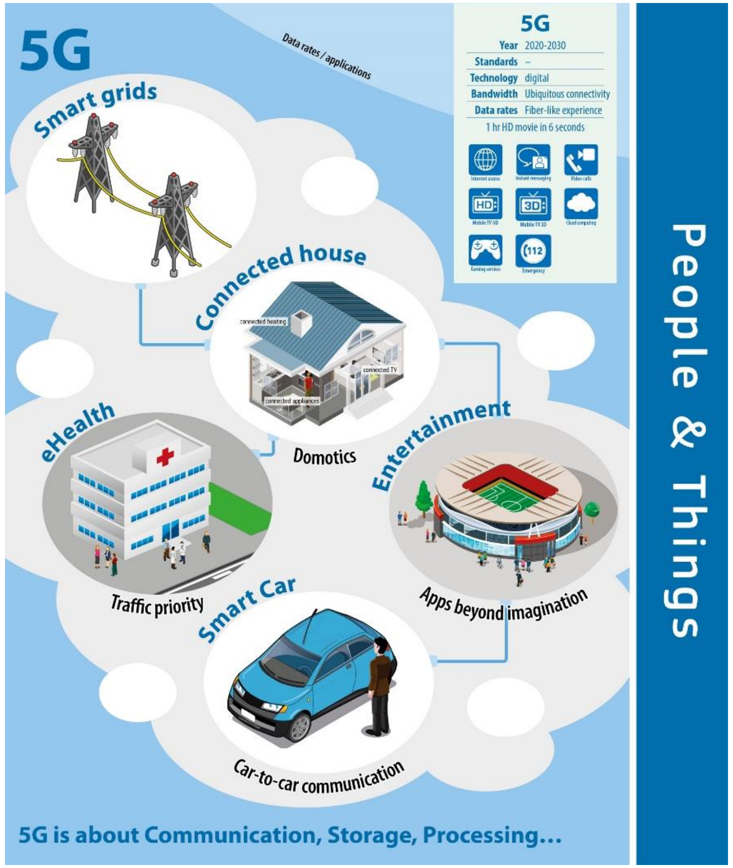
JULY 11- JULY 14 2021

IEEE Computational Intelligence Society

IEEE

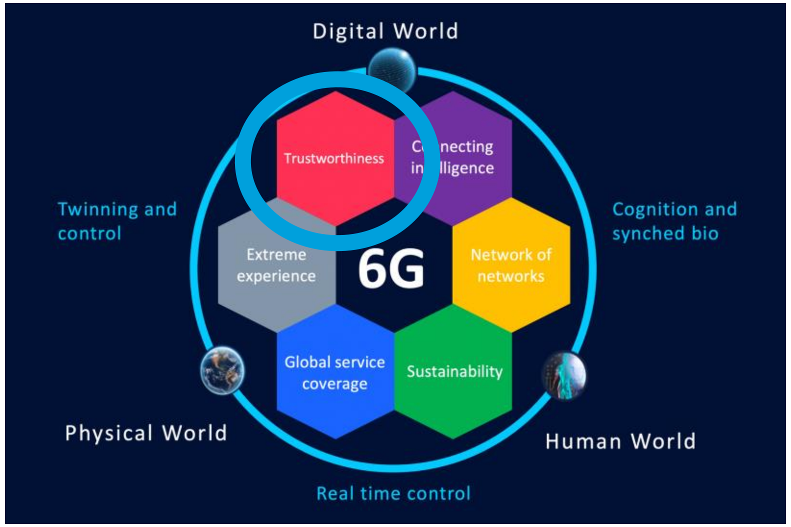
Introduction: AI and Wireless Networks





Figures from https://digital-strategy.ec.europa.eu/en/library/1g-5g-infographic

Figure from Hexa-X Deliverable D1.2, “Expanded 6G vision, use cases and societal values”. Online: https://hexa-x.eu/wp-content/uploads/2021/05/Hexa-X_D1.2.pdf



Crucial role of **AI/ML** techniques

Introduction: Trustworthy AI

Trustworthy AI

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
  - Traceability
  - Explainabilty
  - Communication
- Diversity non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

"AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned."

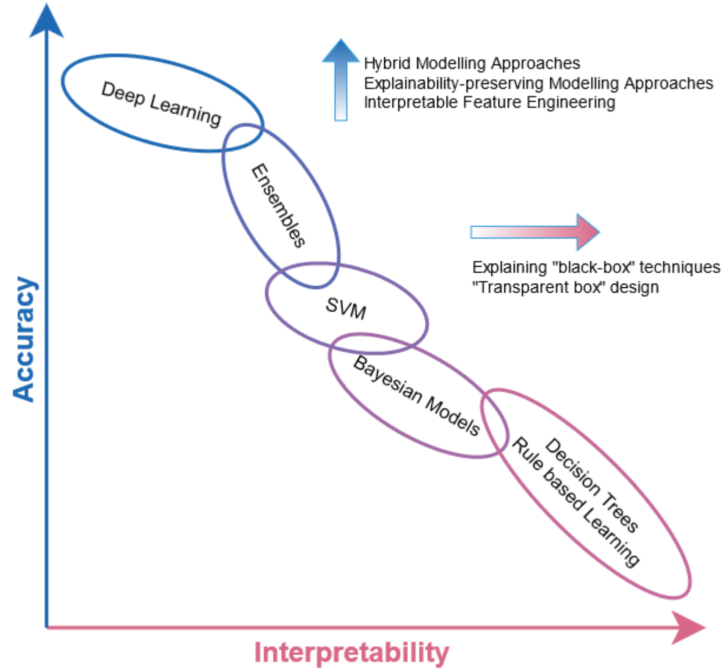
European Commission, *"Ethics guidelines for trustworthy AI"* Report, 2019

Introduction: Trustworthy AI (Cont’d)

Explainabilty

Design of **Inherently Interpretable** Models

***Post-hoc* Explainability** Techniques



Inspired to Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion 58 (2020): 82-115.

Fuzzy Models

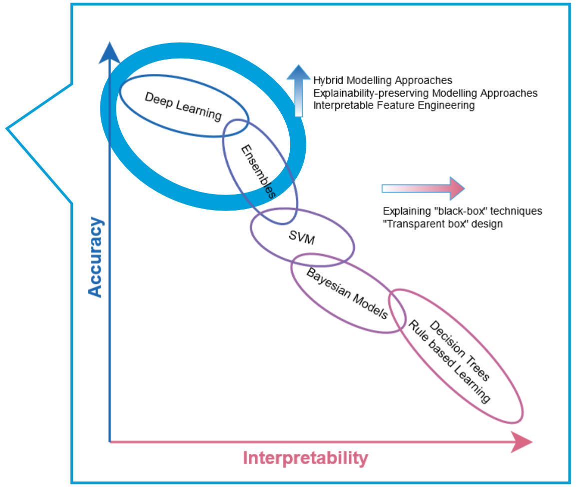
Linguistic representation enhances *semantic* interpretability

Motivation and goals

- Relevant **case study**:
  - Next Generation Networks will have stringent requirements in terms of:
    - Quality of Service (offered by the network)
    - Quality of Experience (user-perceived, tailored on the application)
- Vast majority of **current thrusts** for the adoption of AI for wireless networks are based on **“black-box” models**
- Increasing attention for **Trustworthy AI**

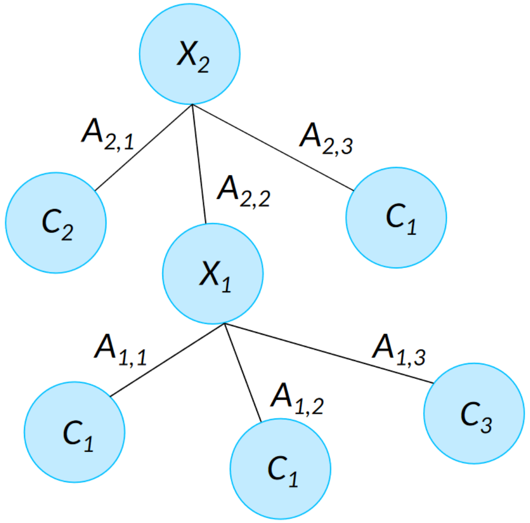
Goals:

- Adoption of **XAI models (Fuzzy Decision Trees)** in wireless networks for **Quality of Experience prediction**
- Experimental comparison: investigate the **explainability/accuracy trade-off** in the context of **tree-based models**

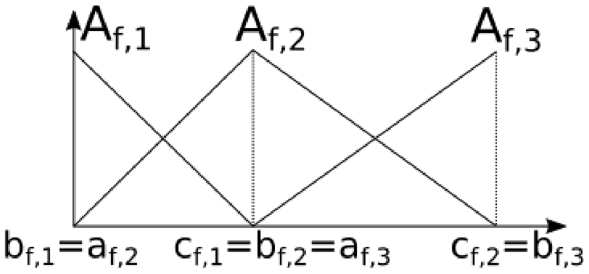


Background: Fuzzy Decision Tree (FDT)

- Directed acyclic graph
- Generated in a top–down way by performing **recursive partitions of the attribute space**.
- Typically, requires a **fuzzy partition defined upon each continuous attribute**.



Multi-way FDT



Strong Fuzzy Partition

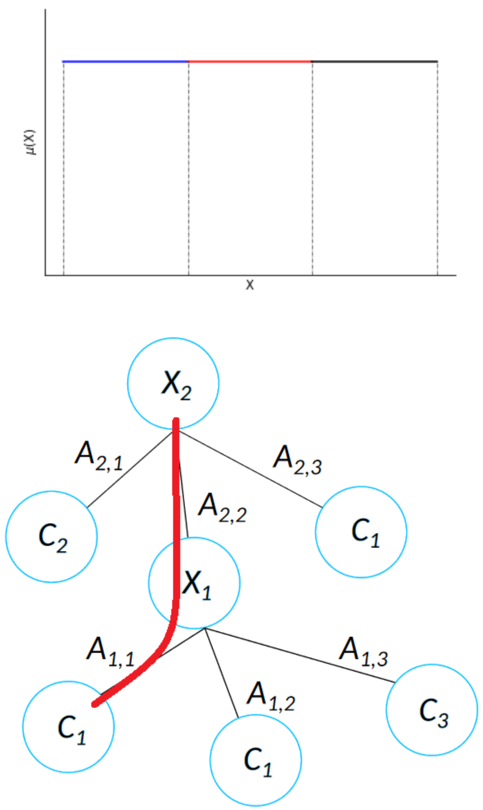
Segatori, Armando, Francesco Marcelloni, and Witold Pedrycz. "On distributed fuzzy decision trees for big data." IEEE Transactions on Fuzzy Systems 26.1 (2017): 174-192.

Background: Fuzzy Decision Tree

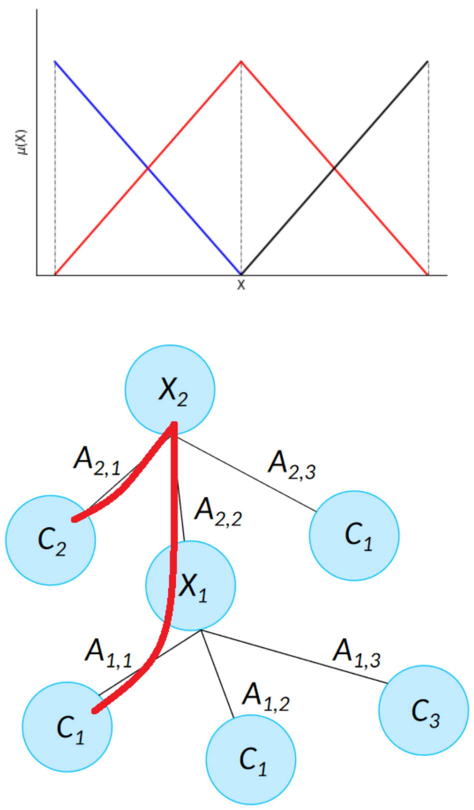
Main factors that affect **explainability** of FDTs

- Structural complexity
  - Numbers of nodes/leaves
- Inference process
  - Maximum *association degree*
- Linguistic fuzzy partition
  - Semantic interpretability
  - *Strong* triangular fuzzy partition

Crisp Decision Tree



Fuzzy Decision Tree



Experimental Setup: *QoS-QoE* Dataset

- Around 69000 **streaming video sessions**
- Goal: to derive a mapping between **QoS metrics** and **QoE factors** (multi-class classification problem)
- Simulated in a fully controllable simulation environment at both *network* and *streaming* levels

29 input variables

TABLE I
INPUT ATTRIBUTES: QOS METRICS AND THEIR DESCRIPTION.

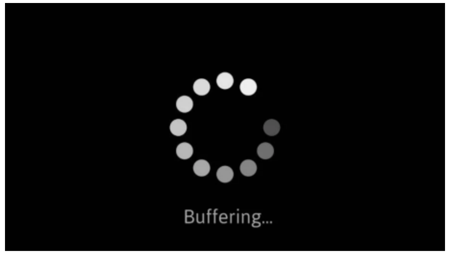
Name	Description
TCP[Output/Input]Packet	Number of TCP packets (In and Out)
TCP[Output/Input]Delay	Avg. delay of TCP packets (In and Out)
TCP[Output/Input]Jitter	Avg. jitter of TCP packets (In and Out)
TCP[Output/Input]Ploss	Loss rate of TCP packets (In and Out)
TCPInputRetrans	Packet retransmissions of TCP
StdNetworkRate	Standard deviation of the network rate
[x]_InputRateVariation; x in {0,5,10,25,50,75,90,95,100}	x^{th} quantile for the network rate (measured in intervals of 2s)
StdInterATimesReq	Std. dev. of inter-arrival times of segment requests
[x]_InterATimesReq; x in {0,5,10,25,50,75,90,95,100}	x^{th} quantile for the inter-arrival times of segment requests

$f(\cdot)$

NoStall [51155: ~74%]

MildStall [17180: ~25%]

SevereStall [794: ~1%]



Vasilev, Vladislav, et al. "Predicting QoE factors with machine learning." 2018 IEEE International Conference on Communications (ICC). IEEE, 2018.

Model evaluation via 5-fold stratified CV

- Model complexity (number of nodes, leaves)
- Evaluation metrics: Precision, Recall, F1-measure per class

Experimental Setup: Classification Models

Model	Multiway Fuzzy Decision Tree	Binary Decision Tree	Random forest (RF)
Implementation	Segatori et al. (2018)	CART scikit-learn	scikit-learn
Depth	**{3,4}** –> {MFDT-3, MFDT-4}	**{6, 11}** –> {BDT-6, BDT-11}	*default*
Splitting criterion	Fuzzy infogain	Infogain	*default*
Partitioning	Fuzzy - *a priori* - supervised	During tree construction	During trees construction
Max fuzzy sets	5	-	-

Experimental Results: Original Dataset

RESULTS ON THE ORIGINAL DATASET. AVERAGE VALUES.

	F1-measure Training	Test
MFDT-3	0.8443	0.8422
MFDT-4	0.8493	0.8472
BDT-6	0.8764	0.8678
BDT-11	**0.9158**	**0.8737**
RF	***0.9999***	***0.8953***

(a) micro-average F1

	Model Complexity Leaves	Nodes
MFDT-3	88.8	***115.2***
MFDT-4	331.6	438.6
BDT-6	***60.2***	119.4
BDT-11	740.0	1479.0
RF	337601.0	675102.0

(b) Model Complexity

	No Stall			Mild Stall			Severe Stall		
	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall
MFDT-3	0.9008	0.8881	0.9149	0.6785	0.7017	0.6639	0.0279	0.1846	0.0151
MFDT-4	0.9032	0.9025	0.9045	0.7015	0.6923	**0.7149**	0.0257	0.1833	0.0138
BDT-6	0.9173	0.9021	**0.9335**	0.7228	0.7588	0.6938	0.5207	**0.7599**	0.3994
BDT-11	**0.9209**	**0.9103**	0.9320	**0.7378**	**0.7658**	0.7142	**0.6068**	0.6548	***0.5681***
RF	***0.9348***	***0.9166***	***0.9539***	***0.7777***	***0.8241***	***0.7376***	***0.6536***	***0.8670***	0.5291

(c) Precision and Recall by class on test set.

- **BDTs** slightly outperform **MFDTs** in terms of micro-average F1-score on the test set
- **RF** achieves highest overall performance but with huge global complexity (and low interpretability)
- *Moderately* low performance on the intermediate class (**Mild Stall**)
- *Considerably* low performance on the minority class (**Severe Stall**)

⇒ **Dataset re-balancing**

Experimental Results: Dataset re-balancing

No Stall: 2000

Mild Stall: 2000

Severe Stall 695

RESULTS AFTER RE-BALANCING THROUGH RANDOM UNDERSAMPLING. AVERAGE VALUES.

	F1-measure	
	Training	Test
MFDT-3	0.7859	0.8119
MFDT-4	0.8167	**0.8335**
BDT-6	0.8508	0.8112
BDT-11	**0.9447**	0.8026
RF	*1.0*	*0.8575*

(a) micro-average F1

	Model Complexity	
	Leaves	Nodes
MFDT-3	115.0	143.6
MFDT-4	396.0	500.8
BDT-6	***55.0***	***109.0***
BDT-11	278.4	555.8
RF	43911.2	87722.4

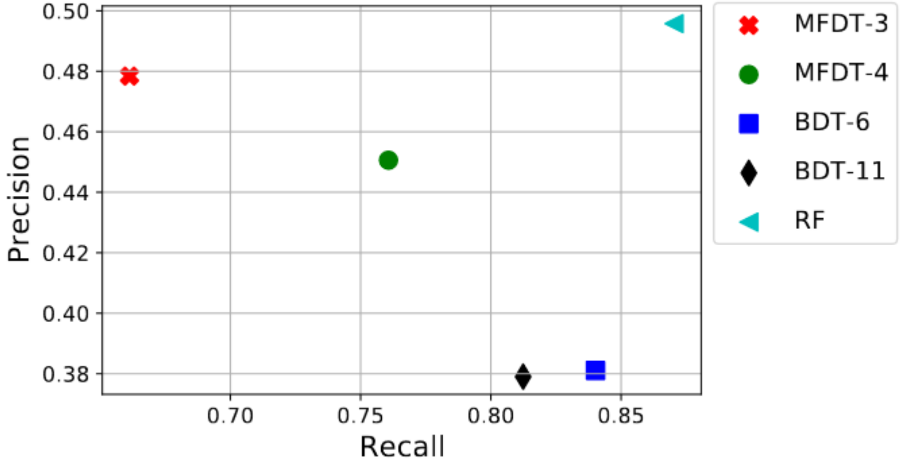
(b) Model Complexity

	No Stall			Mild Stall			Severe Stall		
	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall
MFDT-3	0.8712	0.9396	0.8125	0.6856	0.5916	0.8171	0.5546	**0.4784**	0.6613
MFDT-4	**0.8907**	0.9415	**0.8451**	**0.7071**	**0.6323**	0.8024	**0.5590**	0.4506	0.7607
BDT-6	0.8731	**0.9504**	0.8082	0.6862	0.5920	**0.8189**	0.5235	0.3811	**0.8402**
BDT-11	0.8689	0.9311	0.8146	0.6611	0.5817	0.7665	0.5152	0.3790	0.8124
RF	*0.9070*	*0.9537*	*0.8649*	*0.7452*	*0.6740*	*0.8349*	*0.6291*	*0.4958*	*0.8703*

(c) Precision and Recall by class on test set.

- All models improve their **recall** on **Mild Stall** and **Severe Stall** classes
- **MFDTs** comparable or slightly better than **BDTs** in terms of micro-average F1-score
- **RF** achieves highest overall performance but with huge global complexity (and low interpretability)

Experimental Results: Explainability



(b) Results on class 2: *SevereStall*

$$R_{MFDT-4}: \textbf{IF } \textit{100\_InterATimesReq} \textbf{ is } \textit{VeryHigh}$$
$$\textbf{AND } \textit{25\_InterATimesReq} \textbf{ is } \textit{VeryHigh}$$
$$\textbf{AND } \textit{TCPInputPloss} \textbf{ is } \textit{VeryLow}$$
$$\textbf{THEN } \textit{StallLabel is SevereStall}$$

$$R_{BDT-6}: \textbf{IF } \textit{StdInterATimesReq} > 1.30$$
$$\textbf{AND } \textit{25\_InterATimesReq} > 0.86$$
$$\textbf{AND } \textit{StdInterATimesReq} > 1.59$$
$$\textbf{AND } \textit{25\_InputRateVariation} > 186749.00$$
$$\textbf{AND } \textit{TCPOutputJitter} > 0.00$$
$$\textbf{AND } \textit{90\_InputRateVariation} > 473853.50$$
$$\textbf{THEN } \textit{StallLabel is SevereStall}$$

- BDTs and MFDTs achieve different trade-offs between precision and recall
- BDT-6 outperforms BDT-11 with respect to all objectives

Conclusions

- Adoption of **fuzzy models** for addressing the task of **Quality of Experience classification**
- **Experimental comparison** between tree-based models on a recently proposed **QoS-QoE dataset,** characterized by a severe **class imbalance**
  - **Multiway Fuzzy Decision trees** achieves competitive performance in capturing *stall events,* in terms of precision, recall, micro-avg F1-measure.
  - **Multiway Fuzzy Decision trees** feature higher semantic interpretability than Binary Decision Trees
  - Random Forest outperforms all other models but does not feature inherent interpretability

What's next:

- QoS-QoE as a **time-series prediction** problem
- **Multi-objective evolutionary algorithms** for concurrently optimizing accuracy and complexity of FDTs

Thank you for your attention