Rule-Based Dynamic Feature Selection for Explainable Classification with Uncertainty Quantification

Javier Fumanal-Idocin[1], Raquel Fernandez-Peralta[2,*], Javier Andreu-Perez[1]

[1]School of Computer Science and Electronic Engineering, University of Essex

[2]Mathematical Institute, Slovak Academy of Sciences

*Corresponding author: raquel.fernandez@mat.savba.sk

YUFE4Postdocs

University of Essex

Co-funded by the European Union

Funded by the European Union NextGenerationEU

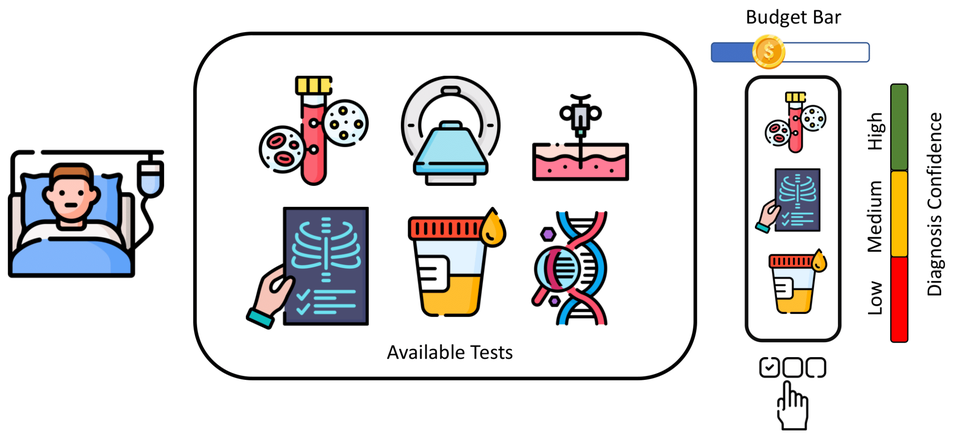
RECOVERY AND RESILIENCE PLAN

DEPUTY PRIME MINISTER'S OFFICE FOR THE RECOVERY PLAN AND THE KNOWLEDGE ECONOMY

SLOVAK ACADEMY OF SCIENCES

What is DFS?

Many real-world decision-making tasks involve high-dimensional data, where obtaining every feature may be costly or impractical (e.g., medical tests). In these settings, it is often beneficial to acquire only the information that is most relevant for each sample.



Dynamic Feature Selection (DFS) addresses this challenge by selecting features sequentially, querying only those expected to provide the greatest value for the prediction.

Existing DFS approaches commonly follow two directions:

- **Reinforcement learning (RL)** methods: flexible and capable of learning acquisition policies, but often computationally demanding and sensitive to training dynamics.
- **Information theory-based greedy approaches**: generally efficient and uncertainty-aware, but reliant on accurate estimation of the underlying measures, e.g., Conditional Mutual Information (CMI).

For random variables X, Y, and Z, the **CMI** $I(X;Y|Z)$ measures the information shared between X and Y given Z. In the DFS context:

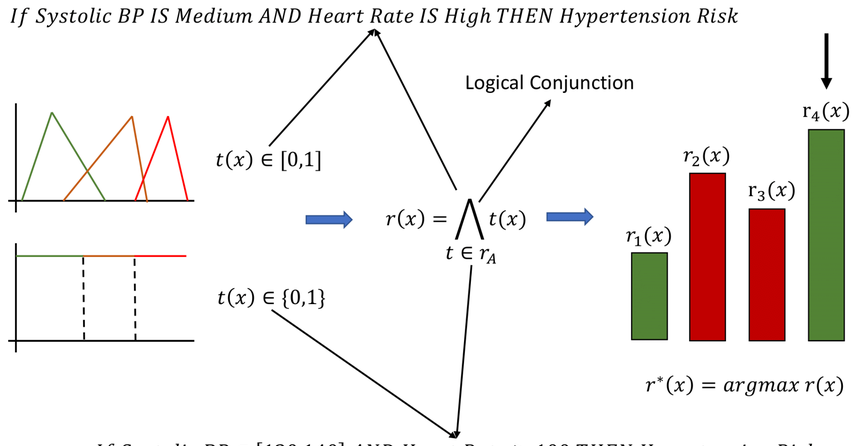
$$I(y;x_i|\mathbf{x}_S) = D_{\mathrm{KL}}(p(x_i,y|\mathbf{x}_S)||p(x_i|\mathbf{x}_S)p(y|\mathbf{x}_S)),$$

where $\mathbf{x}_S$ are the observed features and x_i is a candidate feature.

These approaches often rely on **high-capacity models**, reducing transparency and requiring post-hoc methods to explain feature acquisition and predictions.

Rule-Based Reasoning

Rule-based classifiers use human-readable **IF–THEN** rules linking feature conditions (antecedent) to a predicted class (consequent).



Our contribution: We integrate DFS with **interpretable rule-based models** and **explicit uncertainty quantification**. This enables:

- human-understandable feature acquisition strategies,
- transparent prediction rules,
- interpretable aleatoric and epistemic uncertainty at each step.

DFS Guided by a Rule-Based Global Model

We assume access to a strong global model and our DFS strategy evaluates features by estimating how much adding a variable improves a sub-model's ability to match the global model's predictions captured through the value function v_q that computes the expected quality gain $q(x_i, x_S)$ when including feature x_i:

$$v_q(x_i,\mathbf{x}_S) = \mathbb{E}_{p(x_i|\mathbf{x}_S)}[q(x_i,\mathbf{x}_S)], \quad q(x_i,\mathbf{x}_S) = u(\mathbf{x}_S \cup \{x_i\}) + \lambda e(\mathbf{x}_S \cup \{x_i\}).$$

Prediction Uncertainty: We measure how much a sub-model's prediction differs from the full model by using the KL divergence between their predictive distributions.

$$u(\mathbf{x}_S) = D_{\mathrm{KL}}(p(\hat{y}|\mathbf{x})||p(\hat{y}|\mathbf{x}_S)).$$

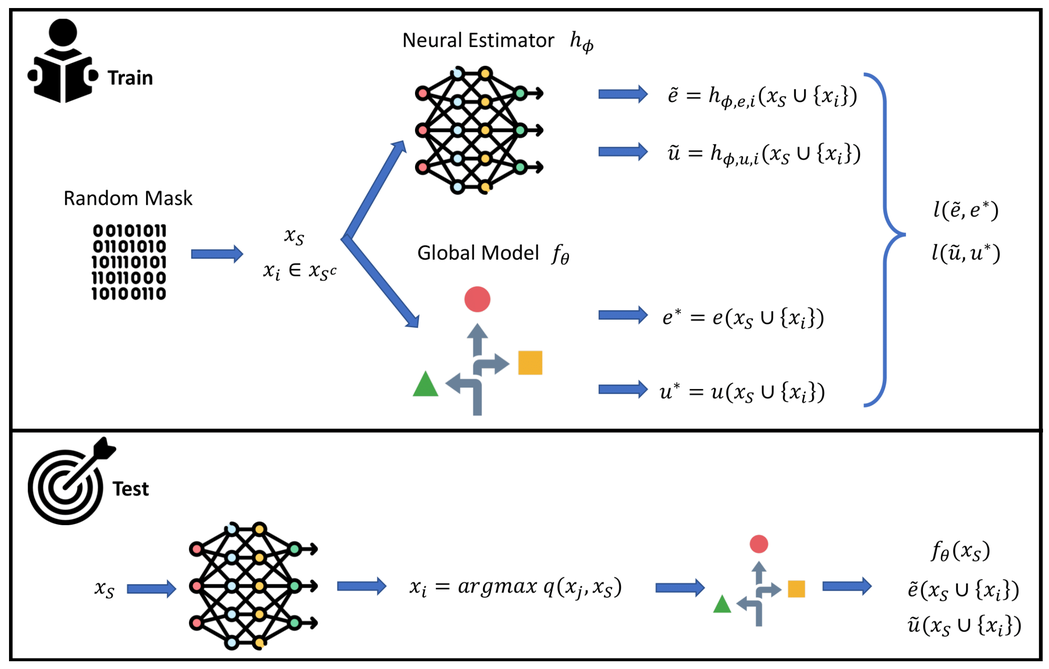
Epistemic Uncertainty concerning rule firings: To evaluate how reliable the winning rule's prediction is, we measure how much it changes across auxiliary models trained on resampled data:

$$e_R(\mathbf{x}_S) = \frac{1}{|I|}\sum_{i\in I} D_{\mathrm{KL}}(p^i_{r^*}(\hat{y} \mid \mathbf{x}_S)||p_{r^*}(\hat{y} \mid \mathbf{x}_S)).$$

For fuzzy rules, we can exploit the fact that a sample can partially activate multiple rules. Each co-firing rule offers an alternative explanation, so we replace auxiliary models with weighted divergences between each rule and the winning rule:

$$e_R(\mathbf{x}_S) = \sum_{r\in R} r(\mathbf{x}) D_{\mathrm{KL}}(p_r(\hat{y} \mid \mathbf{x}_S)||p_{r^*}(\hat{y} \mid \mathbf{x}_S)).$$

Model Training and Evaluation



Experiments and Results

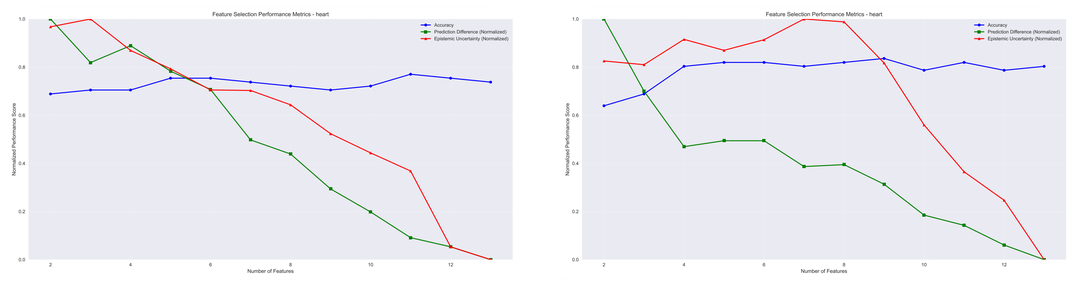


Fig. 1: Comparison of Fuzzy Rule-Based and CART classifier on the Heart dataset. Accuracy (blue), prediction uncert. (green), and epistemic uncert. (red) are shown across varying numbers of features.

	Method	Diabetes	Heart	Cirrhosis	Wine	Yeast	Average (rank)
Static	CMI	64.04 ± 8.10	81.64 ± 11.63	64.60 ± 14.77	95.37 ± 2.54	45.59 ± 4.25	69.25(3.6)
	TabNet	55.20 ± 12.38	80.08 ± 10.22	66.58 ± 5.31	89.81 ± 3.07	53.21 ± 3.56	68.98(3.5)
	CFS	53.27 ± 4.90	71.31 ± 0.43	57.46 ± 3.15	92.22 ± 4.02	48.09 ± 9.37	64.47(6.8)
Dynamic	CWCF	55.78 ± 2.13	78.06 ± 10.19	59.28 ± 7.58	86.66 ± 5.05	45.03 ± 6.01	64.96(6.6)
	INVASE	54.10 ± 7.84	77.36 ± 3.92	61.19 ± 8.26	92.77 ± 1.57	46.03 ± 4.51	66.29(5.6)
	DIME	62.02 ± 1.86	81.75 ± 15.65	66.90 ± 9.04	89.81 ± 2.53	45.59 ± 1.15	69.21(3.6)
	VIP	63.12 ± 0.98	71.20 ± 5.39	66.34 ± 6.52	85.00 ± 3.53	49.53 ± 0.89	67.04(5.2)
	CART-DFS*	55.37 ± 0.32	78.68 ± 10.84	65.95 ± 5.59	95.37 ± 2.01	51.78 ± 3.08	69.43(3.3)
	FuzzyCART-DFS*	53.91 ± 1.03	67.21 ± 10.07	66.40 ± 2.90	91.88 ± 5.72	50.16 ± 5.91	65.91(5.6)

Fig. 2: Comparison of the different feature selection methods tested, both static and dynamic. Results are the average accuracy for the first 10 features queried. * Our proposal.

Conclusions and Future Work

- Rule-based DFS provides a **consistent, interpretable, and uncertainty-aware alternative** to neural and RL-based methods, while remaining competitive.
- **Our method matches the accuracy** of DIME and VIP, and slightly surpasses CWCF and INVASE across benchmarks.
- Strong results from static selectors highlight the **difficulty of designing robust dynamic policies**.

Future work includes joint training of the model and selector, handling non-uniform feature costs, and using DFS to further improve the performance–explainability trade-off.