Adaptive Prototypical Parts in Tabular Data

Jacek Karolczak (jacek.karolczak@doctorate.put.poznan.pl)

Motivation

Tabular data underpins high-stakes decisions in medicine, finance, and justice. Current XAI relies on *post-hoc* methods (SHAP, LIME, counterfactuals) that:

- Demand effortful System 2 reasoning, mismatching natural human cognition [4]
- Risk misrepresenting the true model logic (faithfulness gap)
- Address feature importance and decision boundaries but ignore the data manifold

Cognitive basis

Rosch's categorization theory [14] shows humans reason by comparing to representative exemplars. Prototypical parts align with this fast, intuitive System 1 reasoning.

The Explanation Triad

Effective explanations must address three geometric facets [8]:

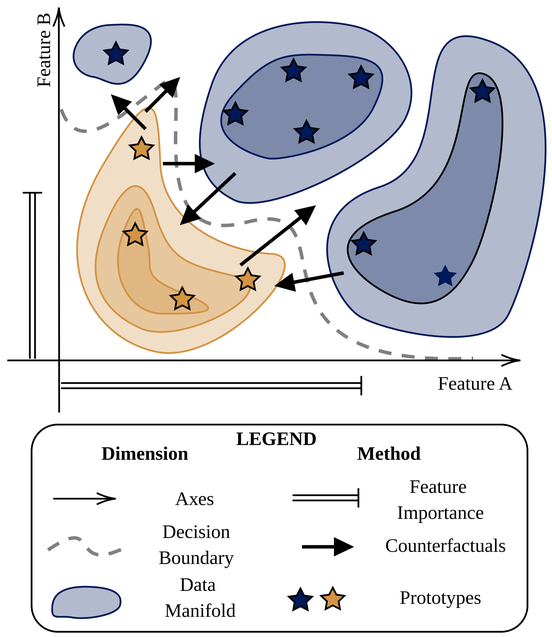


Figure 1: Feature importance defines coordinate axes, counterfactuals probe the decision boundary, prototypes validate the data manifold. From [8].

Prototype-based methods provide the missing link that grounds explanations in actual data.

From Prototypes to Parts

- Standard prototypes compare full instances [5, 10, 12]
- Tabular features are heterogeneous; class instances can be divided into subgroups, which often share only subsets of features
- Alike parts [6]: post-hoc highlighting of shared feature subsets
- Limitation: full-instance representations ⇒ faithfulness issues

Preceding Work: MEDIC

MEDIC [7] expresses prototypical parts as subsets of medical norms (reference intervals). This proves the concept but limits broader applicability.

P^{2}Tab generalizes by:

- Removing domain-specific assumptions
- Using differentiable masking (learned, not prescribed)
- Handling arbitrary mixed-type tabular features

P^{2}Tab: Prototypical Parts for Tabular Data

P^{2}Tab pioneers prototypical parts for tabular data via differentiable patch masking, mapping abstract embeddings to localized feature subsets [9].

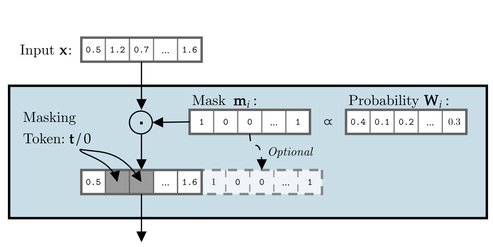


Figure 2: The patching layer decomposes tabular instances into granular feature subsets. From [9].

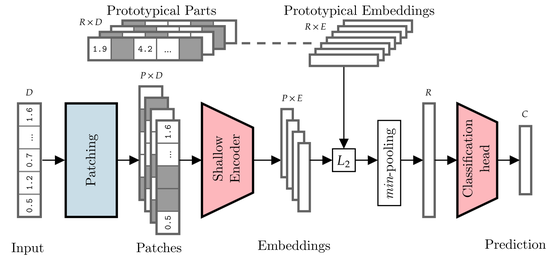


Figure 3: P^{2}Tab: differentiable patching, shared encoding, and distance-based prototype comparison. From [9].

Key properties

- Predictions directly traceable to historical data patterns
- Competitive accuracy with opaque models
- Inherent granular transparency (no post-hoc step)
- Explanations reside within the true data manifold

Transparent Reasoning Trace

	ExternalRiskEstimate	MSinceOldestTradeO.	AverageMInFile	NumTrades60Ever2D.	NumTrades90Ever2D.	PercentTradesNeverD.	MaxDelq2PublicRecL.	NumTotalTrades	NumTradesOpeninL.	MSinceMostRecentI.	NumInqLast6M	NumInqLast6Mexcl.	NetFractionRevol.	NetFractionInstallB.	NumRevolvingTrades.	NumInstallTrades.	NumBank2NatlT.	PercentTradesWB.	...	L_2 Embeddings Distance **d**
Full Instance	88	-8	129	0	0	100	7	10	1	-7	3	3	4	-8	1	1	0	40	...	
Feature Importance	.69	.38	.83	.31	.47	.36	.51	.82	.96	1.0	.36	.84	.47	.36	.58	.16	.40	.47	...	
Prototypical Part 1	84	-	-	-	-	-	6	-	0	-	-	-	-	79	-	-	-	-	-	.0931
Nearest Instance Patch	88	-	-	-	-	-	7	-	1	-	-	-	-	-8	-	-	-	-	-	
Prototypical Part 2	-	-	-	-	0	-	-	36	-	-	-	-	-	-	0	-	-	60	-	.0977
Nearest Instance Patch	-	-	-	-	0	-	-	10	-	-	-	-	-	-	1	-	-	40	-	
Prototypical Part 3	80	-	-	-	-	-	6	-	1	-	-	-	-	-8	-	-	-	-	-	.1270
Nearest Instance Patch	88	-	-	-	-	-	7	-	1	-	-	-	-	-8	-	-	-	-	-	
...									...											...
Prototypical Part 30	-	211	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	9	-	.5202
Nearest Instance Patch	-	-8	-	0	0	-	-	-	-	-	-	-	-	-	-	-	-	0	-	
Prototypical Part 31	-	-	36	-	-	-	-	-	3	-7	-	7	-	-	-	-	-	-	-	.5443
Nearest Instance Patch	-	-	129	-	-	-	-	-	-	-	-	-	4	-	-	1	-	40	-	
Prototypical Part 32	-	-	25	-	-	-	-	-	3	2	-	1	-	-	-	-	-	-	-	.6568
Nearest Instance Patch	-	-	-	-	0	100	-	10	-	-	3	-	-	-	-	-	-	-	-	

$\mathbf{b} + W \exp(-\mathbf{d}) \rightarrow$ Class 1

Figure 4: Final prediction of P^{2}Tab: transparently traced to grounded prototypical parts and their L_2 distances to input patches. From [9].

The linear classification head ensures full traceability: each prediction decomposes into weighted similarity scores to learned prototypical parts.

Predictive Performance

Dataset	Tree	RF	XGB	MLP	P^{2}Tab
CodRNA	.966	.980	**.981**	.972	.967
Credit Card	.658	.652	.655	.653	**.696**
Diabetes	.588	**.712**	**.723**	.644	**.721**
HELOC	.697	.714	**.728**	**.716**	.707
Ionosphere	.810	.907	**.918**	**.973**	**.929**
Pendigits	.888	.937	**.956**	.943	.930
Statlog	**.729**	**.734**	**.729**	.717	**.729**

Table 1: Balanced Accuracy. Bold = not significantly different from the best ($p \geq 0.05$, bootstrap). From [9].

P^{2}Tab is statistically tied with the best on 4/7 datasets and achieves the top result on Credit Card—all with inherent interpretability.

Interpretability Metrics

Dataset	Rank Diversity	Feature Focus
CodRNA	0.139	0.276
Credit Card	0.115	0.179
Diabetes	0.150	0.201
HELOC	0.221	0.260
Ionosphere	0.206	0.162

Table 2: Prototype learning metrics for binary-classification datasets. From [9].

Rank Diversity (upper bound ≈ 0.289): std. dev. of each prototype's relative rank across test instances.

Feature Focus (upper bound = 1.0): Gini coefficient over aggregated feature frequencies in prototypical parts.

Future: Dynamic Self-Adjustment

- Data-driven allocation of prototypes and patches (no manual hyperparameters)
- L_1 regularization on prototype activations for iterative pruning of redundant parts
- Differentiable routing (Gumbel-Softmax [3]) to allocate capacity toward complex decision boundaries

Future: Adaptation to Data Streams

- Extend prototypical parts to evolving distributions with concept drift [2, 11]
- Monitor prototype activation frequency; gracefully forget obsolete concepts
- Integrate drift detectors (ADWIN [1], Page-Hinkley [13]) for targeted architectural updates
- Temporal regularization to maintain explanation stability across time windows

Future: Human-Centric Evaluation

- Structured user studies with domain experts and lay users
- Measure task response time, confidence calibration, error detection
- Evaluate whether prototypical parts reduce cognitive load vs. feature attributions
- Validate constructive distrust: do mismatched prototypes trigger System 2 auditing?

References

1. Bifet & Gavalda, 2007. Learning from time-changing data with adaptive windowing.
2. Gama, 2010. Knowledge Discovery from Data Streams.
3. Jang et al., 2016. Categorical Reparameterization with Gumbel-Softmax.
4. Kahneman, 2011. Thinking, Fast and Slow.
5. Karolczak & Stefanowski, 2024. A-PETE: Adaptive Prototype Explanations of Tree Ensembles.
6. Karolczak & Stefanowski, 2025. This part looks alike this: identifying important parts of explained instances and prototypes
7. Karolczak & Stefanowski, 2025. An Interpretable Prototype Parts-based Neural Network for Medical Tabular Data
8. Karolczak & Stefanowski, 2026. Overlooked and Underutilized: Prototypes are the Missing Piece in Tabular XAI.
9. Karolczak & Stefanowski, 2026. Toward Inherently Interpretable Tabular Models via Prototypical Parts.
10. Kim et al., 2016. Examples are not enough, learn to criticize!
11. Lu et al., 2018. Learning under Concept Drift: A Review.
12. Page, 1954. Continuous Inspection Schemes.
13. Rosch, 1975. Cognitive representations of semantic categories.
14. Tan et al., 2020. Tree Space Prototypes.

Acknowledgments

The author thanks Jerzy Stefanowski for supervision and guidance. This work is supported by Poznan University of Technology under grant 0311/SBAD/0772.