AlignXAI: Aligning Local Explanation Rankings with User-Specific Preferences

Kelem Negasi Amare, Radwa El Shawi Supervisor

Institute of Computer Science, University of Tartu

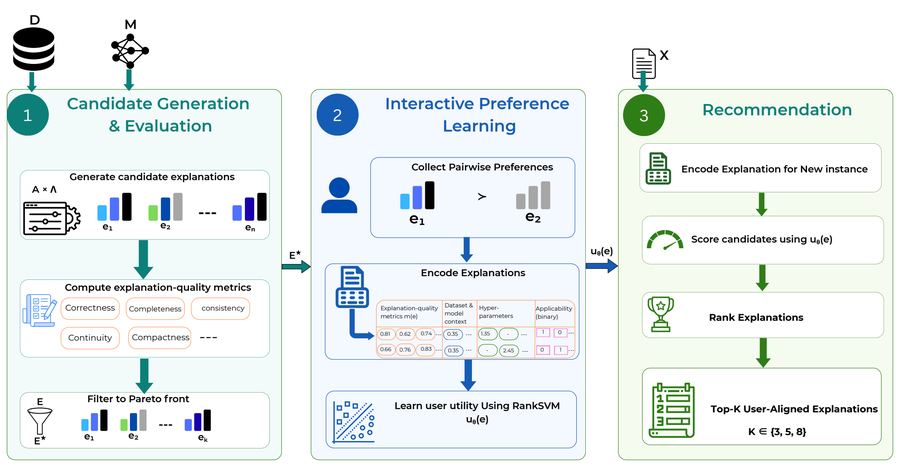

Figure 1. **System overview.** AlignXAI first generates candidate explanations and evaluates them using explanation-quality metrics, then encodes explanations with these metrics and contextual information to learn user-specific preferences from pairwise feedback, and finally recommends top-k user–aligned explanations for new instances.

Introduction

Explainable Artificial Intelligence (XAI) aims to make machine learning predictions more understandable by providing explanations of model behaviour. However, explanations are not only technical outputs; they are also communication tools whose usefulness depends on the user's background, goals, and expertise. For example, a machine learning practitioner, a medical professional, and a novice user may prefer different explanations for the same prediction.

Local XAI methods such as LIME, SHAP, Integrated Gradients, and Causal SHAP generate multiple candidate explanations for the same instance by varying explainer families and their hyperparameters. These candidates may differ in measurable quality properties such as correctness (fidelity), completeness (coverage), consistency (agreement across similar cases), continuity (stability), confidence (reliability), and compactness (simplicity).

Because these quality properties can reflect different trade-offs, a single metric or fixed metric aggregation imposes one predefined notion of what makes an explanation best. AlignXAI instead learns user-specific explanation preferences from pairwise feedback and uses the learned utility model to rank candidate explanations according to how well they align with each user's preferences.

Methodology and Experiments

- **Datasets, Models, and Training:** Experiments use **COMPAS**, **Bank Marketing**, and **German Credit**. For each dataset, four predictive models are trained: Decision Tree (DT), Logistic Regression (LR), Gradient Boosting (GB), and Multilayer Perceptron (MLP).
- **Candidate Generation and Evaluation:** As shown in Figure 1, a trained model M and dataset D are used to generate candidate explanations $E = \{e_1, \ldots, e_n\}$ by varying XAI methods $\mathcal{A}$ and hyperparameters $\lambda \in \Lambda$. Each candidate is scored with explanation-quality metrics $m(e)$ following the Co-12 taxonomy of explanation quality properties introduced by Nauta et al. Pareto filtering then keeps $E^\star \subseteq E$ to preserve diverse trade-offs among explanation properties.
- **Pairwise Preference Simulation:** Present pairs of candidate explanations and record preferences $e_i \succ e_j$. In the experiments, simulated personas define base metric preferences w^0, from which user-specific weights are sampled as $w \sim \mathrm{Dirichlet}(cw^0)$. Each explanation receives latent utility $U(e) = w^\top m(e)$, and noisy pairwise labels are generated with a Bradley–Terry model:

$$P(e_i \succ e_j) = \sigma\left(\frac{U(e_i) - U(e_j)}{\tau}\right),$$

  where c controls persona diversity and τ controls choice noise.
- **Encoding and Utility Learning:** As shown in the second stage of Figure 1, explanations are encoded as

$$\phi(e) = [m(e); z(D); z(M); h(\lambda); a(\lambda)],$$

  combining quality metrics, dataset/model context, hyperparameters, and applicability indicators. SVMRank learns the user-specific utility $u_\theta(e) = \theta^\top \phi(e)$, so preferred explanations receive higher scores.
- **Explanation Search Space:** For each instance, AlignXAI generates 18 candidate explanations by sweeping four XAI methods over the structured hyperparameter grid summarized in Table 1.

Method ($\mathcal{A}$)	Hyperparameters (Λ)	# Candidates
LIME	`num_samples` $\in \{50, 100, 200\}$, `kernel_width` $\in \{1.5, 2.0, 3.0\}$	$3 \times 3 = 9$
SHAP	`bg_size` $\in \{50, 100, 150\}$	3
IG	`steps` $\in \{20, 40, 60\}$	3
Causal SHAP	`coalitions` $\in \{20, 30, 40\}$	3
Total		**18**

Table 1. **Explanation search space.** Hyperparameter grid used to generate candidate explanations for each instance.

Full results and plots

UNIVERSITY OF TARTU
Institute of Computer Science
1632

Results

AlignXAI is compared with AutoXAI, which uses automated hyperparameter optimization and weighted scalarized evaluation based on explicit metric priorities. We evaluate both a full-metrics setting using all explanation-quality indicators (Figure 2) and a three-metric setting using correctness, continuity/stability, and compactness (Figure 3). Across datasets, model families, and $k \in \{3, 5, 8\}$, AlignXAI achieves stronger Spearman alignment with simulated user-utility rankings. The random-preference ablation shows that preference alignment depends on meaningful pairwise feedback rather than random choices (Figure 4), while the comparison-budget sweep shows that most gains are achieved after roughly 10–20 comparisons (Figure 5).

Figure 2. **AlignXAI vs AutoXAI.** Spearman rank correlation between AlignXAI/AutoXAI rankings and simulated user-utility rankings across datasets, model families, and $k \in \{3, 5, 8\}$.

Figure 3. **AlignXAI vs AutoXAI under the three-metric setting.** Spearman rank correlation between explanation rankings and simulated user-utility rankings using correctness, continuity/stability, and compactness across datasets, model families, and $k \in \{3, 5, 8\}$.

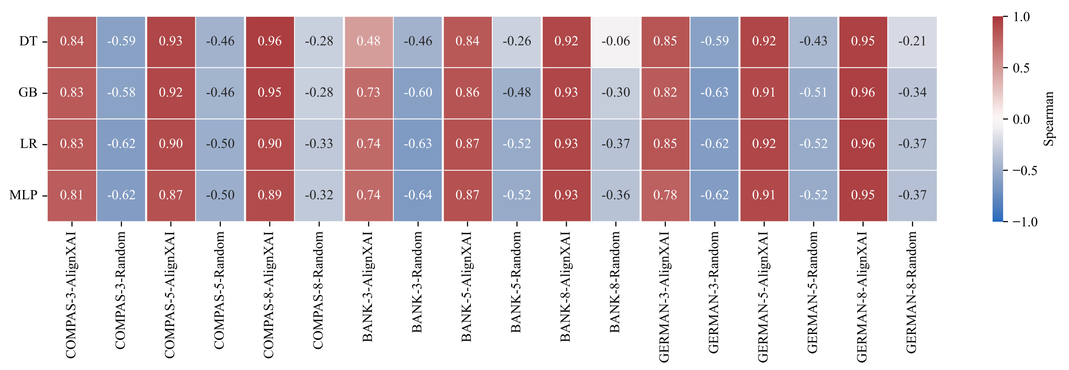

Figure 4. **Ablation study.** AlignXAI is compared with a random pairwise-preference baseline under the same comparison budget. Results report Spearman rank correlation between learned explanation rankings and ground-truth user-utility rankings across datasets, model families, and $k \in \{3, 5, 8\}$.

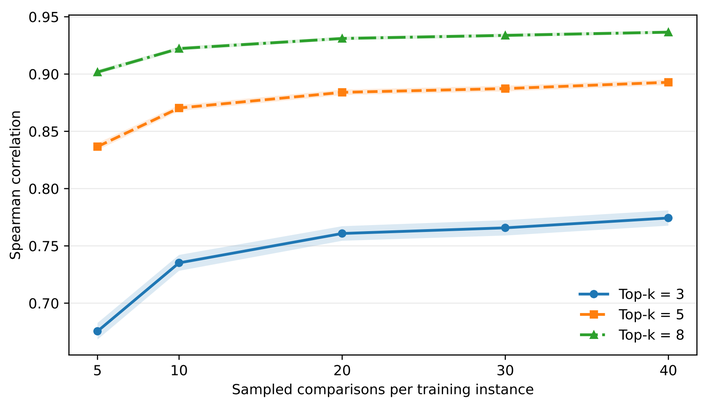

Figure 5. **Effect of pairwise comparison budget.** Spearman rank correlation improves as more pairwise comparisons are sampled per training instance. Most gains occur within the first 10–20 comparisons, after which performance begins to plateau.

Conclusion

AlignXAI aligns local explanation selection with user-specific preferences by learning utilities from pairwise feedback rather than relying on fixed metric aggregation. The results show stronger alignment with simulated user preferences than AutoXAI and random-preference baselines, while also showing that reliable performance can be achieved with only a small amount of user feedback. Overall, this demonstrates that interactive preference learning can make XAI recommendations more user-aligned, adaptive, and sensitive to different explanation needs.

Implementation Code

This work is submitted to CIKM 2026.