

Integrating Feature Understandability into Feature Selection for Human-Centred Tabular Explanations

Nicola Rossberg^{1,2}[0009–0005–3883–5833], Bennett Kleinberg⁴[0000–0003–1658–9086],
Barry O’Sullivan^{1,2,3}[0000–0002–0090–2085], Luca Longo^{1,2,3}[0000–0002–2718–5426], and
Andrea Visentin^{1,2,3}[0000–0003–3702–4826]

¹ School of Computer Science and Information Technology, University College Cork, Ireland

² Centre for Research Training in Artificial Intelligence, University College Cork, Ireland
n.rossberg@cs.ucc.ie

³ Insight Centre for Data Analytics, University College Cork, Ireland

⁴ Department of Methodology and Statistics, Tilburg University

Abstract. Existing human-centred XAI research predominantly evaluates explanations after they have been generated. However, this fails to provide actionable insights into improving accessibility. This study incorporates human-based feature understandability into feature selection to improve the generation of human-centred explanations. Natural language explanations for machine learning models trained on tabular data often cite influential features in the input data to explain a given decision. As a consequence, users’ ability to understand the produced explanation hinges on their understanding of the cited features. To quantify this natural-language feature understandability, Rossberg et al. [53] proposed the Feature Understandability Scale, which assigns numerical ‘understandability scores’ to features based on user feedback. To evaluate the degree to which these scores can improve machine learning explanation quality, this study collects understandability scores across two datasets, proposes an understandability-aware feature selection algorithm to select features that are both understandable and informative, and trains a random forest based on these feature subsets. The SHAP-based explanations generated by this model are evaluated using user feedback, and changes in explanation quality are assessed. The individual evaluation of the explanations showed that while the understandability-based explanations scored higher on average, the difference was not statistically significant. In contrast, the comparative evaluation found that participants preferred these explanations significantly more frequently than expected at the chance level. It is concluded that the integration of understandability can yield more understandable explanations at minimal cost to classification performance, although changes to quality require elucidation in larger cohorts.

Keywords: Psychometrics · Machine Learning · Interpretability by Design · Explainable Artificial Intelligence · Evaluation Methods

1 Introduction

In recent years, the use of machine learning has become increasingly pervasive across a wide range of domains, including impactful areas such as medicine and agriculture

[50, 48, 8, 59]. However, while black-box models are often powerful in predictive performance, they may lack transparency, which is essential, especially in critical fields such as medicine and finance [51]. The ability to audit machine learning systems is crucial to prevent the deployment of inequitable models that result from biased data or poor training. To promote the auditability of black-box models, a wide range of explainable AI (XAI) techniques has been proposed [68]. Through the implementation of such techniques, unstable decisions, especially in situations of uncertainty [52], or those made on the basis of irrelevant data, can be identified and potentially mitigated, thereby supporting the legal compliance and safety of machine learning models [43].

While the implementation of XAI techniques is thus important to build user trust and prevent model bias, explaining is a social process and thus requires consideration of the end-user [46]. Current techniques to measure explanation quality hinge on post-hoc evaluation, building on user feedback or algorithmic computations to quantify the accessibility of the explanation [62, 67, 30]. While this provides insight into the quality of the explanation, it does not show which aspects of the explanation need to be changed to improve the usability. For textual explanations, in which the model’s inferences are converted and expressed in natural language [65], references to input features are one actionable lever to improve explanation quality, with user understanding hinging on their comprehension of the features included in the explanation. To thus create an actionable explanation evaluation, it is necessary to quantify understanding of the feature, rather than the explanation level.

To this end, Rossberg et al. [53] proposed the *Feature Understandability Scale* (FUS), which measures feature understandability across various dimensions using user input. An example of the scale’s implementation is provided in Table 8 in the Appendix. In the final use case, the scale is distributed to a representative sample of the audience receiving the final explanation. The understandability scores (US) assigned to each feature are computed by taking the average user rating. The US can then be incorporated into the machine learning pipeline through various modalities, including cost-sensitive feature selection or a Pareto frontier. By favouring more understandable features during model training, the explanations produced by the machine learning model are expected to be more accessible, thereby increasing their utility. This is illustrated in Figure 1

This study aims to fulfil three successive objectives, which are outlined below:

1. Collect a representative sample of understandability ratings across two datasets and analyse their distributions;
2. Propose an understandability-aware feature selection mechanism, and train a machine learning model on the basis of the selected feature subsets;
3. Evaluate the quality of the explanations generated on the basis of the trained machine learning model.

The rest of this report is structured as follows. Section 2 provides an overview of the relevant literature pertaining to XAI, cost-sensitive feature selection and the explainability-accuracy trade-off. Section 3 describes the steps taken to collect understandability ratings, build the feature selection algorithm and evaluate the generated explanations. Section 4 then presents and analyses the results of each of the three steps of the study. Finally, Section 5 summarises and evaluates the results and Section 6 details the limitations and future research directions of this study.

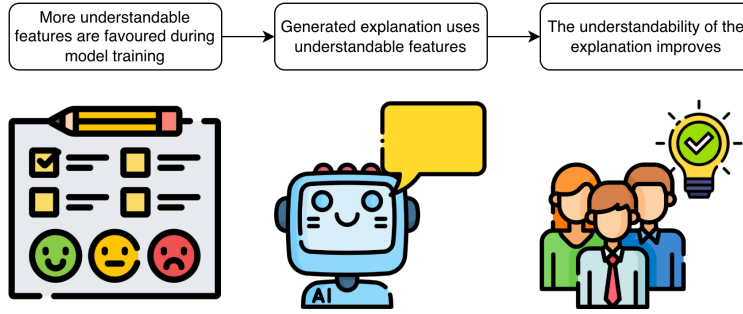


Fig. 1: Graphical depiction of the theoretical basis on which the FUS implementation is expected to increase the understandability of explanations for tabular data.

2 Literature Review

2.1 Explainable Artificial Intelligence

While the pervasiveness of Machine Learning models has increased across a range of domains in recent years, the dangers associated with their black-box nature, including bias and shortcut learning, have been significant barriers to their widespread implementation [10, 42, 61, 57]. XAI techniques address these concerns by auditing decision processes and promoting transparency and trust [69, 27]. A large number of XAI techniques have been proposed, including feature attribution [56, 37], counterfactuals [38, 9, 29], latent-space mechanistic interpretability [2, 74, 35, 1], and natural language methods [21, 75, 15]. A full review and taxonomy of XAI methods can be found at Vilone and Longo [66].

2.2 Interpretability by Design

One important aspect considered in recent works is the construction of models that are interpretable by design, integrating transparency into the model’s architecture and training process [54, 25]. Examples of such models may include learning of predictors through interpretable queries [18], the inclusion of human judgment in the model design process [39], and the integration of human feature evaluations [76, 64]. While interpretability by design promotes faithful explanations for machine learning systems, it has been argued that it may lead to lower performance, as transparent algorithms lack the necessary complexity [5, 31, 60, 33, 4]. However, while this is theoretically sound, it is seldom observed in practice, and it is therefore argued that more transparent models should be favoured by default [12, 24, 17].

2.3 Cost-Sensitive Feature Selection

Cost-sensitive feature selection incorporates cost into the feature selection framework, where cost is either the test cost or the misclassification cost of including the feature

[13, 41, 14]. Test costs refer to the cost of feature acquisition, for example, through medical tests, where a more expensive test may yield a more accurate result [32]. Costly misclassifications are domain-specific and user-specified. An example is the misclassification of cancerous cells as benign, which is more costly to public health than misclassifying healthy cells as cancerous, and features may be selected to promote false positives over false negatives [73].

In the literature, three types of cost-sensitive feature selection methods are distinguished: filter methods, wrapper methods and embedded methods [71]. This work focuses on filter methods, as they are the most transparent and straightforward to implement. Filter methods pre-select features before model training, using feature cost as a scoring function. Features are then selected based on a user-defined threshold or feature number, selecting only the N cheapest features. Filter methods are advantageous as they are computationally cheap and can be applied to high-dimensional datasets [6].

This work differs from the traditional literature on cost-sensitive feature selection in that feature costs are user-defined rather than error- or test-based, and success is evaluated based on explanation quality rather than misclassification patterns. Where in the traditional literature the cost of a feature is defined either by its contribution to a costly misclassification or through the cost of its acquisition, this work defines a costly feature as one that is not understandable and thus contributes to ineffective explanations [41, 73, 14]. Unlike traditional literature, success is assessed by the final understandability of explanations and the accuracy of the machine learning model.

2.4 SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) is based on cooperative game theory [58] and was among the earliest XAI techniques to provide both global (whole-model) and local (instance-specific) explanations. The implementation of SHAP values to measure feature importance was first proposed by [40, 55] and later expanded for applications to Neural Networks [36], feature selection [20] and outcome explanation [63]. However, the widespread use of SHAP began only when [45] published their unified framework, which combined previous approaches to create an XAI method that provides both local and global explanations, with model-agnostic and model-specific modalities [55].

2.5 Explanation Evaluation

The evaluation of explanation quality has been investigated in past literature, and several different approaches have been proposed [72, 7, 70, 47, 23, 22]. The properties of a good explanation are user and context-dependent. In general, it is important that the explanation be on the right level and include sufficient detail for the targeted user group [34]. The length of the explanation should also be tailored to the usage scenario; where a lay-user may have less time to read an explanation, a practitioner or data scientist may be willing to spend the time parsing a longer explanation [19]. When decomposing explanations to their purpose, Halpern et al. [26] propose that a good explanation provides an answer to a 'why' question. In addition, Hoffman et al. [28] state that a good explanation permits an a priori judgement regarding its quality and hence must be both precise and clear. While these metrics permit intermittent evaluation of

explanations through qualitative analysis, the final assessment of an explanation should be based on user feedback.

The collection of user-based evaluations of explanations should be reproducible and consistent. In the field of psychometrics, this is achieved through the development and validation of scales that reliably capture user attitudes [16]. Vilone and Longo [67] proposed a 12-item scale to measure the quality of machine-generated rule-based explanations, in which the user is queried about the clarity they have gained from the explanations. In a similar vein, Holzinger et al. [30] proposed the system causability scale, which measures the quality of a human-AI interface or explanation process. This allows the user to assess the extent to which the communication method makes the provided explanations accessible.

2.6 Synthesis and Literature Gaps

This study aims to fill several gaps identified in the existing literature. To the best of the authors’ knowledge, there are currently no studies integrating self-reported feature understanding into the machine learning training pipeline on an a priori basis. Second, the study expands on existing cost-sensitive feature selection methods by introducing a new approach to cost quantification and model evaluation. Finally, the study proposes two statistically rigorous explanation evaluation procedures, based on user preference and individual, scale-based evaluations.

3 Method and Design

The methodology of this research is divided into three key steps. First, the FUS is applied across two datasets, and the US for each feature are computed. Second, two feature selection approaches are implemented in parallel: the first implements the understandability-sensitive feature selection, identifying features that are both informative and understandable, referred to as the co-optimised setting going forward. The second selects features solely based on informativeness without considering understandability, and is henceforth referred to as the informative setting. For each approach and dataset, a random forest model is implemented, optimised and evaluated. In the final step, SHAP is used to extract feature importance from the trained machine learning models, and a heuristic is applied to generate textual explanations. The explanations are evaluated through user feedback, and differences in explanation quality between the co-optimised and informative settings are analysed. The code for the complete analysis is available on GitHub⁵

3.1 Step 1 - Understandability Score Collection

The FUS is applied to two public datasets to ensure that any changes in model explanation quality are not domain-specific. An example of the implementation of the FUS is shown in Table 8 in the Appendix, and the full scales can be found in [53]. The first dataset is the ‘Telco Customer Churn’ dataset, which contains data that can be

⁵ github.com/ncrossberg/Explanation-Evaluation-Paper.

used to predict whether a customer will switch service providers given their family and usage history. The dataset is publicly available ⁶. This dataset will be referred to as the ‘telco dataset’ going forward. The second dataset is the ‘Body Signals of Smoking’ dataset and contains data useful for the prediction of whether a patient is a smoker, given a range of physiological indicators. The dataset is publicly available ⁷ and will be referred to as the ‘medical dataset’ going forward. Before collecting the US, the datasets were cleaned by removing features with zero variance. Furthermore, user ID variables were removed because they were unique patient identifiers and lacked predictive relevance. No other features were removed during preprocessing. The final telco dataset consisted of 19 features and 7043 instances, and the medical dataset consisted of 24 features and 55,692 instances. It is important to note that while diverse datasets are chosen to promote theoretical generalisability, practical transferability of findings to other domains cannot be guaranteed.

US of the remaining features after preprocessing were collected via a Qualtrics⁸-based survey, distributed through Prolific⁹. All data collection was approved by the University College Cork Ethics Committee for Social Research under Log Number 2025-231A4. Participants had to be over 18 and be nationals of, and reside in, the United States of America, the United Kingdom, or Ireland to be eligible. All participants had to provide informed consent before commencing the study.

Participants were randomly assigned to one dataset and asked to rate half of the features using the FUS. Each participant evaluated a total of 10 features (telco dataset) or 12 (medical dataset). Participants were asked to rate only half the features to limit workload and avoid respondent fatigue. Features were presented in a pseudo-random order and counterbalanced to ensure that each feature received an approximately equal number of ratings. Feature-wise attention checks were implemented, and features with failed checks were removed. Finally, the mean US were computed for each feature. The features and respective mean and total US can be found on GitHub.

3.2 Step 2 - Feature Selection and Machine Learning Model Implementation

To ensure that the reported performance results of the final models were not due to the chosen train-test split, a repeated Monte Carlo sampling procedure was implemented prior to feature selection and machine learning model training to assess the equivalence of different data splits. The data were split into a stratified 70:30 train-test split across 50 different seeds to assess the distribution of model performance. The mean, standard deviation and normal-approximation confidence intervals were computed across the fifty splits.

Subsequently, the data were split into training and testing sets in a 70:30 ratio, and the subsequent procedures were implemented on the training set only. Figure 2 shows the full workflow of this step.

⁶ <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

⁷ <https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking>

⁸ Qualtrics is a survey-building and distribution website <https://www.qualtrics.com/>

⁹ Prolific connects researchers with participants, enabling large-scale data collection: <https://www.prolific.com/>

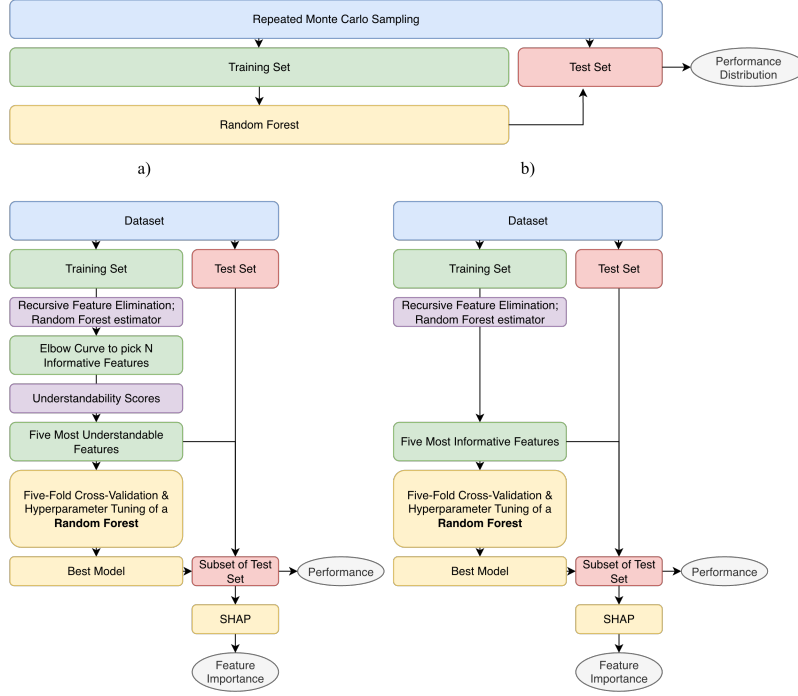


Fig. 2: Visualisation of the Repeated Monte Carlo Sampling, a) co-optimisation and b) informative setting of the feature selection, and machine learning model training methodology.

Feature Selection Two feature selection procedures were implemented in parallel, for the co-optimised and informative setting, respectively. Feature informativeness for both settings was computed through recursive feature elimination (RFE) with a random forest estimator. In the co-optimised setting, feature importance was plotted, and the top n features were identified using an elbow function. Of this subset of size n , the five most understandable features were retained as the final feature subset. In the informative setting, the top five most important features were selected directly from the fitted RFE model. A final subset of size five was selected to constrain explanation length and limit cognitive load, while still providing sufficient information to the user.

Machine Learning Model Implementation After the feature subset generation, a random forest classifier was trained and optimised. A random forest was chosen as the model because it is less prone to overfitting than decision trees, can efficiently handle categorical data, and is well-suited to large datasets. Hyperparameter tuning was implemented using five-fold cross-validation on the training set, and the best-performing model was retrained on the entire training set and evaluated on the final test set. To generate the final explanations, ten instances were selected from the test set, and TreeSHAP [44] was implemented to extract feature importance values.

3.3 Step 3 - Explanation Evaluation

Based on the feature importances extracted in step 2, the final textual explanations were generated using the heuristic outlined below:

1. Compute instance i's feature contributions through Tree SHAP
2. Partition features into two sets by contribution direction;
 - SUPPORT = features increasing the predicted outcome
 - DETRACT = features decreasing it
3. Verbalise each feature:
 - numeric $\rightarrow$ compare value to its mean ("above/below average")
4. Compose explanation:
 - State prediction, justified by SUPPORT features (in rank order)
 - If DETRACT is non-empty, append as a caveat clause
5. Return explanation string.

Explanations were corrected for grammar using the software "Grammarly" to ensure cohesion and legibility. The generated explanations were evaluated through user feedback to assess differences in quality between the co-optimised and informative settings. Qualtrics was used to implement the evaluation procedure, and participants were recruited via Prolific. Data was collected on the 15th of September 2026, and participants were remunerated monetarily in line with Irish minimum wage requirements. To be eligible, participants had to be over the age of 18 and nationals of and residing in Ireland, the United Kingdom or the United States. They also could not have previously completed a survey related to this research. All participants had to digitally sign an informed consent form prior to commencing the survey.¹⁰

The explanation evaluation procedure was split into two distinct approaches: individual and comparative evaluation, the details of which are outlined below. A total of ten prototype explanations were selected for each dataset and explanation setting (40 explanations total). The explanations are expected to function as a prototype of the optimisation procedure, and a mixed effects model was implemented to assess the prototype-wise variation. Participants were randomly assigned to one of the evaluation procedures. The full survey workflow is outlined in Figure 3.

Individual Evaluation The individual evaluation used an adapted version of the scale developed by Vilone & Longo [67]. The questionnaire was adapted by removing two items: "No rules in the explanatory method return surprising predictions" and "The explanatory method only includes the most relevant variables from the data". The items were excluded, as explanations in this study are not rule-based, and the training set consists of only five features, precluding participants from assessing whether these are the most relevant, as they do not have access to the full feature set prior to selection. The remaining items were adapted by replacing 'explanatory method' with 'explanation'. The final scale is shown in Table 7 in the Appendix.

During the evaluation, participants were presented with two pairs of two explanations (Co-optimised and informative) for each dataset and asked to rate the explanations

¹⁰ The informed consent form can be found on Github github.com/ncrossberg/Explanation-Evaluation-Paper

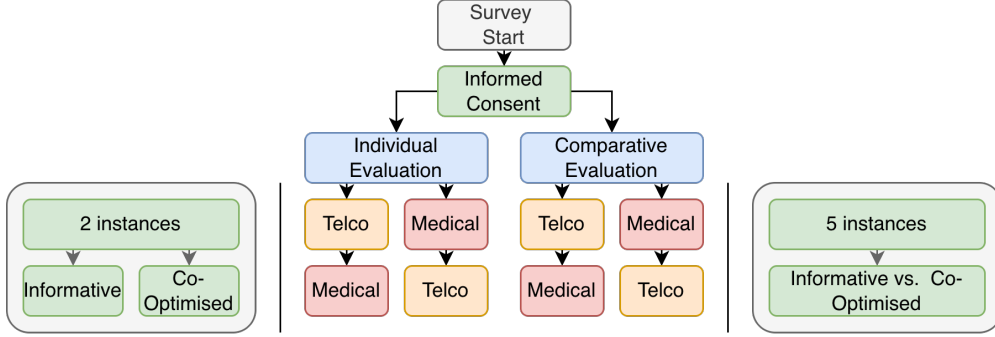


Fig. 3: The Survey Flow of the evaluation procedure. Participants were randomly assigned to the comparative or individual evaluation, and the order of datasets was randomised for each condition. The individual setting used the explanation quality scale [67] to evaluate informative and co-optimised explanations separately, whereas the comparative setting prompted participants to select the more understandable explanation and provided prototypes of both co-optimised and informative explanations. Two or five instances were selected, respectively, and participants were asked to either use the scale to evaluate them (individual evaluation) or select the more understandable explanation (comparative evaluation).

using the scale. The order of datasets and explanation type was randomised and counterbalanced to prevent order effects. After data collection, responses were cleaned, relevant scale items were reverse-coded, and total scores were computed for each prototype and instance. The survey was structured as a repeated-measures, paired design, with participants evaluating multiple explanations across both datasets. To account for the random error introduced by participants, a mixed-effects model is implemented. Each dataset is analysed separately, with the explanation type used as a fixed effect to predict the total score, while accounting for variation between participants and prototypes by including them as random intercepts.

Comparative Evaluation The comparative evaluation setting presented participants with two explanations for the same instance, generated by the co-optimised and informative settings, respectively. Participants were asked to select the explanation they understood better. Participants were presented with five pairs of explanations for each dataset (twenty explanations total). Presentation order of datasets and explanations was randomised and counterbalanced to prevent order effects.

The analysis of the comparative results was conducted in two steps. First, descriptive statistics were computed, providing the total and question-wise preference count for each dataset to assess the general trends in explanation understandability. Subsequently, a mixed-effects logistic regression model was implemented to assess the significance of the number of preferences while accounting for participant and prototype effects. Models were implemented separately for each dataset, and preference was used as the outcome variable, with a random intercept included for both the participant and prototype. The models were implemented using the ‘lme4’ library [11].

4 Results and Discussion

4.1 Feature Rating Collection

Feature ratings were collected on the 31st of August 2026. The average number of ratings per feature was 30.32 (minimum = 22, maximum = 37). Figure 4 shows the distribution of mean US across both datasets, along with the standard errors for each feature, where higher scores indicate greater understandability. The range of US is limited across the two datasets (3.52-4.52 for the telco dataset and 2.94-4.00 for the medical dataset). Several hypotheses can be put forward for this behaviour. While it may be a genuine reflection of limited difficulty in the datasets, it may also be due to social desirability (users reporting inflated understanding to appear better) or overconfidence (users overestimating their own understanding). However, further testing is required to confirm these stipulations.

Interestingly, the telco dataset is considered more understandable on average, possibly due to the absence of specialised medical terms. However, cross-dataset comparisons should be treated with caution, since participants were only exposed to a single dataset, and they may have developed internal benchmarks based on the specific features they encountered. This lack of a shared baseline means that subjective interpretations of the rating scale were not standardised across the two groups, making direct comparisons of the US unreliable. Interestingly, features accessible at face value, such as ‘hearing (right)’, were assigned comparatively low understandability, which may be attributable to the measurement units, and future interview studies may wish to investigate the factors leading to users’ rating decisions.

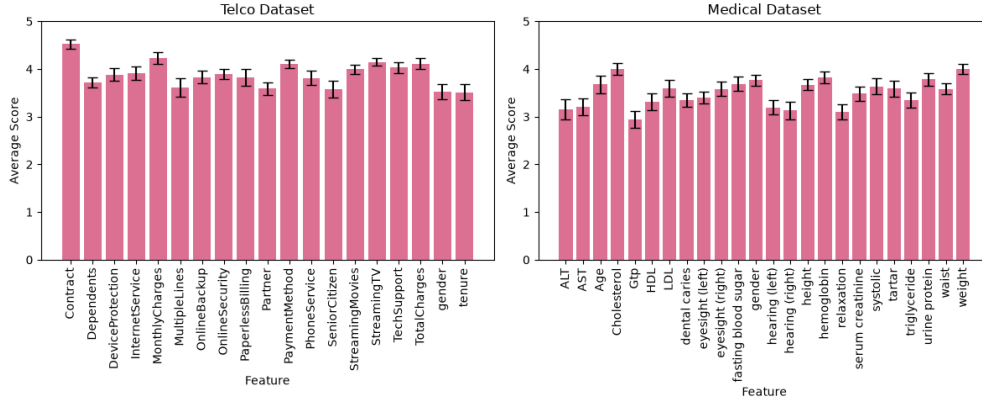


Fig. 4: Distribution of US across the two datasets

4.2 Co-Optimisation

Monte Carlo Resampling The Monte Carlo resampling was conducted across fifty seeds, implementing a stratified 70:30 train-test split and training and evaluating a

random forest on the full feature set for each split. The random forest model was implemented through Scikit-learn, and with the exception of balanced class weights, all default parameters were maintained, to isolate the effect of the split from that of hyperparameter tuning [49]. Table 1 reports the means, standard deviations and confidence intervals (computed through normal approximation) for the resampling procedure.

This highlights that the variation in performance across seeds is limited, and the seed selection for the splitting procedures does not materially affect the reported results. Seed 42 was therefore maintained throughout both feature selection and modelling settings.

Table 1: Distributional Data of the Repeated Monte Carlo Sampling for both Datasets

Telco Data					
		Mean	SD	Lower 95% CI	Higher 95% CI
Balanced Accuracy		0.733	0.009	0.731	0.736
F1		0.603	0.013	0.600	0.607
Precision		0.569	0.013	0.565	0.573
Recall		0.642	0.020	0.637	0.648
Medical Data					
Balanced Accuracy		0.831	0.003	0.830	0.832
F1		0.781	0.003	0.780	0.782
Precision		0.709	0.004	0.708	0.710
Recall		0.869	0.005	0.867	0.870

Feature Selection The performance results of the co-optimisation procedure are shown in Table 2. Notably, the performance of the random forest is stable across the co-optimised and informative settings. The largest observed difference is in the Telco dataset’s recall, with the informative setting performing approximately 2% better. Overall, the classifier’s performance is stable across both settings, indicating that selecting features that are both understandable and informative has negligible effects on performance. This is furthermore supported by the changes in performance lying within or close to the standard deviation observed in the Monte Carlo sampling in the telco dataset, showing that the performance gap between settings is within the noise floor established prior. Importantly, the US of the selected features in the co-optimised setting is higher than that of the informative setting across both datasets. While the difference is rather small (0.20 and 0.25), it represents changes of 19.86% and 22.63% of the total US range in the two datasets, respectively. As such, the model increases the understandability by approximately 20% of the observed range (5% of the total range) without affecting classification performance.

The selected features across the two datasets are presented in Table 3. During initial feature selection in the co-optimised setting, nine features were retained for the telco

Table 2: Results of the Co-Optimisation Procedure for both datasets

	Telco Dataset		Medical Dataset	
	Co-Optimised	Informative	Co-Optimised	Informative
Balanced Accuracy	75.25	75.35	75.54	75.40
F1	61.11	61.03	70.21	70.09
Precision	49.29	48.48	55.64	55.40
Recall	80.39	82.35	95.11	95.37
Mean US	4.16	3.96	3.79	3.54

dataset and eight for the medical dataset, using the elbow of the importance plot to determine the total cut-off. Of the five most understandable features retained, two overlap with the informative setting in the telco dataset and three in the medical dataset. At face value, the features selected by the co-optimised setting are more understandable, giving preference to variables which are more commonly encountered in everyday life, especially in the medical dataset.

Table 3: Features selected for each dataset in the co-optimised and informative setting. The features are ranked by importance as determined by SHAP. Features selected by both settings are bolded.

Telco Dataset		Medical Dataset	
Co-Optimised	Informative	Co-Optimised	Informative
Internet Service, Tech Support , Total Charges, Monthly Charges, Contract	Tenure, Online Backup, Online Security, Tech Support , Contract	height(cm) , age, gender , haemoglobin , weight(kg)	Gtp, serum creatinine, height(cm) , gender , haemoglobin

Explanation Generation A total of 40 prototype explanations are generated across the four categories (two settings and two datasets). An example of one of the explanations is presented below:

The individual is predicted to change companies because their Contract is Month-to-month, their Internet Service is Fibre optic, their Tech Support is no, and their Monthly Charges of 96.05 are above average. The fact that their Total Charges of 1740.70 are less than average detracted from this prediction.

4.3 Evaluation

126 participants were recruited for the evaluation procedure. Participant ages ranged from 19 to 81, with a mean age of 42.76 (SD = 13.90); 61.1% female. Each participant

was assigned to either the individual or comparative evaluation and rated either 2 or 5 pairs of explanations per dataset, respectively. The responses were split by evaluation type and dataset. Before analysis, non-English speakers were removed from the data, removing three participants, and response times were then examined. Response times were processed separately for each evaluation, accounting for differences in task type and resulting response speed. A box-and-whiskers plot of response time was examined, and outliers were removed from the data, leading to the deletion of four participants (one participant from the individual and three from the comparative setting). The final datasets comprised 236 data points from 59 participants for the individual evaluation and 300 data points from 60 participants for the comparative setting, respectively.

Individual Scale-based Analysis Exploratory data analysis was then conducted. The distribution of total scores across the datasets and the two explanation types is shown in Figure 5. The co-optimised explanations score higher on average, although the difference is small ($\Delta_{Telco} = 1.03$ and $\Delta_{medical} = 0.08$). Given the repeated-measures design of the survey, a mixed-effects model is necessary to confirm whether the difference is statistically meaningful.

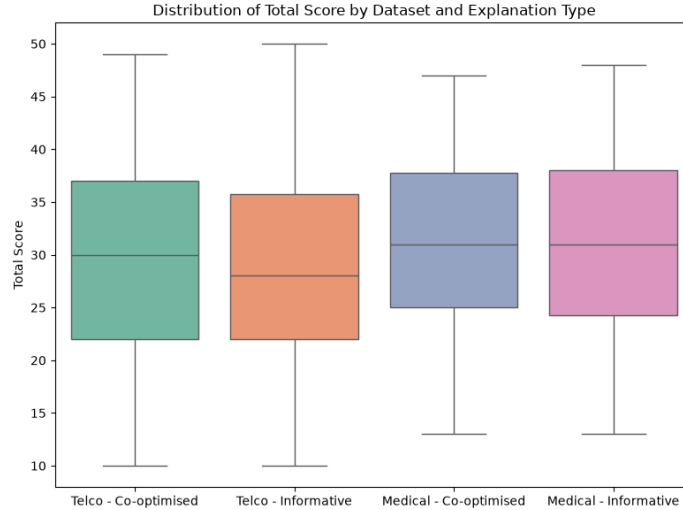


Fig. 5: Distribution of Scale ratings across the two datasets and settings.

The two mixed effects models (one for each Dataset) were implemented using the ‘lme4’ library in R [11]. The p-values were estimated using Satterthwaite’s method, and the confidence intervals are based on likelihood profile computations. The results for each model are presented in Table 4. After the initial model specification, the assumptions of linearity, homoscedasticity, normality of residuals, normality of random effects and singularity were assessed and found to be satisfied.

The mixed-effects model predicted Total Score (range: 10-50) with random intercepts for Participants and Question prototype to account for repeated measures across participants (n=59) and questions (n=10) in both datasets. The model was implemented using the lme4 package in R [11]. The results of both analyses are shown in Table 4. Explanation type had no significant effect on total score, with the informative setting scoring lower than the co-optimised setting by an estimated 1.03 points (Telco) and 0.08 points (medical). While the quality change is in the expected direction, it is not sufficiently strong to reach statistical significance. Especially in the medical domain, the change is minimal and may be an artefact of sample fluctuations. While these results may genuinely reflect a lack of change in quality between explanations, they may also be an artefact of the study’s small sample size, which lacks sufficient statistical power to reject the null hypothesis. Alternatively, the overlap in selected features, especially in the medical dataset, may lead to insufficient differentiation between explanations, consistent with the observed relative change in quality. Finally, the evaluation scale has not been psychometrically validated and hence may not permit the reliable evaluation of quality change.

The Intraclass Correlation Coefficient (ICC) in Table 4 shows that of the variance not explained by explanation type, 55.4% (telco) and 79.7% (medical) is explained by systematic differences between participants. This shows that some participants consistently rate explanations higher or lower than others, regardless of the explanation type. The reasons impacting the discrepancy in judgment between participants should be elucidated in future research and considered as an additional predictor. The example accounted for only 6.7% and 0.9% of the variance, indicating that the impact of the explanation prototype on evaluation was limited.

Table 4: Results of the mixed effects model analysis for the two datasets

Telco Dataset						Medical Dataset				
Fixed Effects	Estimate	SE	95% CI	t (df)	p	Estimate	SE	95% CI	t (df)	p
(Intercept)	29.55	1.31	[26.95, 32.14]	22.64 (29.20)	< .001	31.16	0.99	[29.20, 33.12]	31.33 (61.86)	< .001
Exp. Type [Info]	−1.03	0.75	[−2.51, 0.45]	−1.38 (164.41)	.171	−0.08	0.45	[−0.97, 0.80]	−0.19 (167.18)	.851
Random Effects				SD (Coefficients)						
SD (Int: Participant)			6.99					7.01		
SD (Int: Example)			2.43					0.74		
SD (Residual)			5.78					3.46		
Model Diagnostics					ICC					
Participant			0.554					0.797		
Example			0.067					0.009		
Adjusted ICC			0.621					0.806		

Comparative Analysis Two steps were taken to analyse the forced-choice results. First, the total and per-item number of preferences per dataset was computed. All results are shown in Table 5. Subsequently, a mixed-effects logistic regression was implemented to test the statistical significance of the observed trends while accounting for participant and item-wise variation in preference.

Descriptive Statistics Overall, the co-optimised methodology performed better across both datasets ($\Delta = 70$ in the telco dataset and $\Delta = 100$ in the medical dataset). At the question level, the co-optimised explanations performed better on the telco dataset, with only one example (Q6) showing a higher number of preferences for the informative explanation. In the medical dataset, the responses were more mixed, with two examples (Q5 and Q7) showing marginally better performance for the informative explanation. The differences between the two datasets may be linked to their respective domains, with participants likely being less familiar with medical terminology or preferring more technical terms in medical explanations, which may indicate the model’s proficiency. This is consistent with the trend observed in the individual evaluation. However, this remains conjecture and requires systematic confirmation. It is important to note that, despite best efforts to counterbalance the question presentation, there is some variation in how many times each example was presented.

Table 5: Results of the forced-choice analysis with the preferred option for each question in bold.

Telco Dataset	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Sum	Mean	SD
Informative	13	10	9	8	11	17	15	11	9	12	115	11.5	2.84
Co-Optimised	18	20	21	23	17	12	17	19	19	19	185	18.5	2.92
Medical Dataset	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Sum	Mean	SD
Informative	5	13	7	4	15	12	17	9	9	9	100	10	4.22
Co-Optimised	26	18	23	27	11	19	15	20	20	21	200	20	4.78

Statistical Analysis The results of the mixed-effects logistic regression are shown in Table 6. Choice was specified as the outcome variable, with random intercepts for participant and explanation prototype. The latter was found to have zero variance in the telco dataset and subsequently removed. Overall, the co-optimised explanation was found to be preferred significantly more frequently than expected at chance level, while accounting for participant (and prototype) wise variation. The assumptions of the model were checked using simulation-based scaled residuals, tests of uniformity, outliers and normality of residuals. Influential participants were tested for using leave-one-out refits. All assumptions were found to be satisfied, and tests indicated a good fit.

In the telco dataset, participants selected the co-optimised explanation significantly more often than chance (log-odds = 0.6, $SE = 0.20$, $z = 3.05$, $p = 0.002$; $OR = 1.82$, $95\%CI[1.24, 2.67]$), which corresponds to an estimated 64.5% probability of preferring the co-optimised explanation. In the medical dataset, a larger effect in favour of the co-optimised explanations was found (log-odds = 0.92, $SE = 0.28$, $z = 3.32$, $p < 0.001$; $OR = 2.52$, $95\%CI[1.46, 4.34]$), corresponding to an estimated 71.5% probability of preferring the co-optimised explanation.

Across the two datasets, participant-level variability was similar ($ICC(telco) = 0.261$, $ICC(medical) = 0.275$), indicating that individual differences in baseline pref-

erence accounted for approximately one quarter of the latent-scale variance for both datasets. Prototype-level variability was small in the medical dataset ($ICC = 0.063$) and zero in the telco dataset, suggesting that the preference for the co-optimised explanations was mostly consistent across the presented instances.

Table 6: Results of the mixed-effects logistic regression for the comparative evaluation (outcome: co-optimised explanation chosen = 1).

	Telco Dataset					Medical Dataset				
Fixed Effects	Estimate	SE	95% CI	<i>z</i>	<i>p</i>	Estimate	SE	95% CI	<i>z</i>	<i>p</i>
(Intercept)	0.60	0.20	[0.22, 1.02]	3.05	.002	0.92	0.28	[0.36, 1.55]	3.32	< .001
Derived Estimates	Back-transformed from the intercept									
Odds ratio [95% CI]	1.82 [1.24, 2.67]					2.52 [1.46, 4.34]				
<i>P</i> (co-optimised) [95% CI]	64.5% [55.3, 72.8]					71.6% [59.3, 81.3]				
Random Effects	SD (Coefficients)									
SD (Int: Participant)	1.08					1.17				
SD (Int: Prototype)	-					0.56				
Model Diagnostics	ICC (latent scale)									
Participant	0.261					0.275				
Prototype	-					0.063				
Observations	300					300				
Participants	60					60				

Note: The prototype random intercept was omitted from the telco model due to a singular fit (variance estimated at zero).

Across both evaluation settings, the co-optimised explanation was found to perform better than its informative counterpart. While the individual explanation did not show a significant change in explanation quality, the co-optimised setting yielded directionally higher evaluations. The comparative setting, on the other hand, showed that the co-optimised explanations were selected as more understandable at a level significantly above chance.

5 Conclusion

Overall, this study aimed to deploy the FUS to collect US across two datasets, propose and implement an understandability-aware feature selection mechanism and evaluate the quality change of the explanations built on the machine learning model trained on the feature subset. The collected US showed a limited range across both datasets, which may be attributed to a genuine lack of variability in feature understandability but may also be a result of social desirability or a malfunctioning of the scale. Further validation in other domains is necessary to confirm this behaviour. The performance of the random forest remained essentially unchanged between the two settings, while the

mean understandability score was successfully increased. While the individual evaluation did not show a significant change in explanation quality, the results were directionally in favour of the co-optimised setting. This was supported by the comparative evaluation, which demonstrated that the co-optimised explanations were considered to be more understandable at a level significantly higher than chance. Given that the forced choice is a more sensitive paired comparison, in contrast to the evaluation scale, which is unvalidated and may be noisier, these results strongly indicate that the integration of understandability into the feature selection process can produce more understandable explanations without degrading model performance.

6 Limitations and Future Work

The current study took a filter approach to the co-optimisation of understandability and accuracy. While this permitted the maintenance of high accuracy levels and the promotion of understandable features at low computational cost, future research may wish to ascertain whether wrapper or embedded methods can do so more effectively. In addition, the current study treats understandability as a spectrum. However, it may be of interest to discretise the understandability values into 'levels' to ascertain how this affects explanation quality. Furthermore, the current comparative setting requested participants to select explanations they understood better, where future research may wish to measure objective comprehension, rather than subjective understanding.

In addition to implementing other cost-sensitive feature selection methods, future research may consider more complex machine learning models. This study tested a Random Forest for classification, but it may be of interest to assess whether the accuracy-understandability trade-off can be further minimised with more complex models such as Gradient Boosting or small neural networks designed for tabular data, such as TabNet [3]. While these networks pose some implicit challenges, such as reduced transparency, integrating understandability and employing ad hoc explainability methods may help address them.

The final explanations were evaluated both individually and comparatively. While the informative explanations were found to perform worse than their co-optimised counterparts, the difference in the individual evaluation was not statistically significant and, especially in the medical dataset, was sufficiently small to potentially be due to sampling noise. While this may reflect a genuine lack of quality difference, the comparative evaluation suggests that the co-optimised explanations are generally understood better, which may be due to the small sample size or overlap in selected features between the two settings. Future research may thus wish to repeat this across different domains and with larger cohorts to confirm trends and assess the extent to which results are reproducible.

Several limitations to the evaluation procedure need to be acknowledged. The current evaluation procedure relied on relatively small sample sizes (236 - 300 observations from 59-60 participants per dataset), limiting the generalisability of the results. Additionally, the scale used to evaluate the quality of explanations has not been formally validated. While the items align with the current study's goal, the validity of the ratings cannot be guaranteed. Based on this, an important avenue for future research may be the psychometric validation of a scale evaluating the quality of explanations. Furthermore,

while this study showed some quantitative changes in the quality of the explanations, qualitative interviews may be useful for identifying other areas for improvement in the structure and composition of the explanations, as well as the impact of the quantity and presence of specialised (non-understandable) terms.

Acknowledgments. This work was conducted with the financial support of Research Ireland - Taighde Éireann, under Grant Nos. 18/CRT/6223 and 12/RC/2289-P2, the latter being co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix

Appendix

Table 7: Scale adapted from Vilone & Longo used for the individual explanation evaluation.

Number	Facet	Item
1	Actionability	I have learned something from the explanation
2	Causality rev.	The relationship between the input data and the predictions is vague
3	Comprehensibility rev.	The structure of the explanation is not clear
4	Efficiency rev.	It took me a long time to understand the explanation
5	Explicitness rev.	The explanation is incomprehensible
6	Informativeness	The explanation provides useful information
7	Intelligibility rev.	External support was required to understand the explanation e.g., books, internet search, another person
8	Interestingness rev.	The explanation is not interesting
9	Mental fit	The explanation allows me to understand the logic of the machine learning model used to generate the predictions
10	Security	Thanks to the explanation, I believe that the model will return accurate predictions for all reasonable inputs

Table 8: “Please rate the feature ‘Credit Score’ using the following questions. The scores of ‘Credit Score’ range from 18.4 to 118.65. The feature is used to predict whether or not a customer’s loan application will be approved. While rating the feature, consider if you would understand a loan decision, if it was justified using the feature. E.g. ‘The customer did not get a loan because their Credit Score was 56.95’.”

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I can understand the scale (units) of the feature.	x				
I can easily understand if a given value of the feature is high or low.				x	
I know what this feature measures.				x	
I know what this feature represents.		x			
I think that I can easily access a definition of the feature.					x
In my opinion, the feature should be used to predict the outcome.			x		
I think that the feature is important for the outcome.					x
I think it is fair that the feature influences the outcome.				x	

Bibliography

- [1] Ahmed, T., Biecek, P., Longo, L.: Latent space interpretation and mechanistic clipping of subject-specific variational autoencoders of eeg topographic maps for artefacts reduction. In: World Conference on Explainable Artificial Intelligence. pp. 327–350. Springer (2025)
- [2] Ahmed, T., Longo, L.: Examining the size of the latent space of convolutional variational autoencoders trained with spectral topographic maps of eeg frequency bands. *IEEE Access* **10**, 107575–107586 (2022)
- [3] Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence. pp. 6679–6687. No. 8 (2021)
- [4] Assis, A., Dantas, J., Andrade, E.: The performance-interpretability trade-off: A comparative study of machine learning models. *Journal of Reliable Intelligent Environments* **11**(1), 1 (2025)
- [5] Assis, A., Vêras, D., Andrade, E.: Explainable artificial intelligence-an analysis of the trade-offs between performance and explainability. In: 2023 IEEE Latin American Conference on Computational Intelligence (LA-CCI). pp. 1–6. IEEE (2023)
- [6] Bach, M., Werner, A.: Cost-sensitive feature selection for class imbalance problem. In: International Conference on Information Systems Architecture and Technology. pp. 182–194. Springer (2017)
- [7] Balog, K., Radlinski, F.: Measuring recommendation explanation quality: The conflicting goals of explanations. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. pp. 329–338 (2020)
- [8] Bao, Z., Bufton, J., Hickman, R.J., Aspuru-Guzik, A., Bannigan, P., Allen, C.: Revolutionizing drug formulation development: The increasing impact of machine learning. *Advanced Drug Delivery Reviews* **202**, 115108 (2023)
- [9] Baron, S.: Explainable ai and causal understanding: Counterfactual approaches considered. *Minds and Machines* **33**(2), 347–377 (2023)
- [10] Bassi, P.R., Cavalli, A., Decherchi, S.: Explanation is all you need in distillation: Mitigating bias and shortcut learning. *arXiv preprint arXiv:2407.09788* (2024)
- [11] Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. *Journal of statistical software* **67**, 1–48 (2015)
- [12] Bell, A., Solano-Kamaiko, I., Nov, O., Stoyanovich, J.: It’s just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. pp. 248–266 (2022)
- [13] Benítez-Peña, S., Blanquero, R., Carrizosa, E., Ramírez-Cobo, P.: Cost-sensitive feature selection for support vector machines. *Computers & Operations Research* **106**, 169–178 (2019)
- [14] Bian, J., Peng, X.g., Wang, Y., Zhang, H.: An efficient cost-sensitive feature selection using chaos genetic algorithm for class imbalance problem. *Mathematical Problems in Engineering* **2016**(1), 8752181 (2016)

- [15] Cahlik, V., Alves, R., Kordik, P.: Reasoning-grounded natural language explanations for language models. In: Guidotti, R., Schmid, U., Longo, L. (eds.) *Explainable Artificial Intelligence*. pp. 3–18. Springer Nature Switzerland, Cham (2026)
- [16] Carpenter, S.: Ten steps in scale development and reporting: A guide for researchers. *Communication methods and measures* **12**(1), 25–44 (2018)
- [17] Carrington, A., Fieguth, P., Chen, H.: Measures of model interpretability for model selection. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. pp. 329–349. Springer (2018)
- [18] Chattopadhyay, A., Slocum, S., Haeffele, B.D., Vidal, R., Geman, D.: Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(6), 7430–7443 (2022)
- [19] Chen, Z., Subhash, V., Havasi, M., Pan, W., Doshi-Velez, F.: What makes a good explanation?: A harmonized view of properties of explanations. *arXiv preprint arXiv:2211.05667* (2022)
- [20] Cohen, S., Dror, G., Ruppín, E.: Feature selection via coalitional game theory. *Neural computation* **19**(7), 1939–1961 (2007)
- [21] Danilevsky, M., Dhanorkar, S., Li, Y., Popa, L., Qian, K., Xu, A.: Explainability for natural language processing. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & data mining*. pp. 4033–4034 (2021)
- [22] Dieber, J., Kirrane, S.: A novel model usability evaluation framework (muse) for explainable artificial intelligence. *Information Fusion* **81**, 143–153 (2022)
- [23] Domnich, M., Veski, R.M., Välja, J., Tulver, K., Vicente, R.: Predicting satisfaction of counterfactual explanations from human ratings of explanatory qualities. In: *World Conference on Explainable Artificial Intelligence*. pp. 210–229. Springer (2025)
- [24] Dziugaite, G.K., Ben-David, S., Roy, D.M.: Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764* (2020)
- [25] Gao, Y., Meng, Q., Zhou, Y., Pan, L.: Towards intrinsic interpretability of large language models: A survey of design principles and architectures. In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 34741–34754 (2026)
- [26] Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science* (2005)
- [27] Heider, M., Albrecht, M., Schilp, J., Hähner, J.: Assessing stakeholder perspectives on the explainability of ai solutions for smart production planning with just-in-time logistics. *Expert Systems with Applications* p. 131352 (2026)
- [28] Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018)
- [29] Höllig, J., Markus, A.F., De Slegte, J., Bagave, P.: Semantic meaningfulness: evaluating counterfactual approaches for real-world plausibility and feasibility. In: *World Conference on Explainable Artificial Intelligence*. pp. 636–659. Springer (2023)
- [30] Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations. *KI-Künstliche Intelligenz* **34**(2), 193–198 (2020)

- [31] Hunsicker, T., König, C.J., Langer, M.: Investigating choices regarding the accuracy-transparency trade-off of ai-based systems across contexts. *Computers in Human Behavior: Artificial Humans* p. 100216 (2025)
- [32] Jiang, L., Kong, G., Li, C.: Wrapper framework for test-cost-sensitive feature selection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **51**(3), 1747–1756 (2019)
- [33] Kabir, S., Hossain, M.S., Andersson, K.: A review of explainable artificial intelligence from the perspectives of challenges and opportunities. *Algorithms* **18**(9), 556 (2025)
- [34] Kaplan, S., Uusitalo, H., Lensu, L.: A unified and practical user-centric framework for explainable artificial intelligence. *Knowledge-Based Systems* **283**, 111107 (2024)
- [35] Kästner, L., Crook, B.: Explaining ai through mechanistic interpretability. *European journal for philosophy of science* **14**(4), 52 (2024)
- [36] Keinan, A., Sandbank, B., Hilgetag, C.C., Meilijson, I., Ruppin, E.: Fair attribution of functional contribution in artificial and biological networks. *Neural computation* **16**(9), 1887–1915 (2004)
- [37] Koenen, N., Wright, M.N.: Toward understanding the disagreement problem in neural network feature attribution. In: *World Conference on Explainable Artificial Intelligence*. pp. 247–269. Springer (2024)
- [38] Kuhl, U., Artelt, A., Hammer, B.: For better or worse: the impact of counterfactual explanations’ directionality on user behavior in xai. In: *World Conference on Explainable Artificial Intelligence*. pp. 280–300. Springer (2023)
- [39] Lage, I., Ross, A., Gershman, S.J., Kim, B., Doshi-Velez, F.: Human-in-the-loop interpretability prior. *Advances in neural information processing systems* **31** (2018)
- [40] Lipovetsky, S., Conklin, M.: Analysis of regression in game theory approach. *Applied stochastic models in business and industry* **17**(4), 319–330 (2001)
- [41] Liu, M., Xu, C., Luo, Y., Xu, C., Wen, Y., Tao, D.: Cost-sensitive feature selection by optimizing f-measures. *IEEE Transactions on Image Processing* **27**(3), 1323–1335 (2017)
- [42] Liu, Y., Liu, C., Zheng, J., Xu, C., Wang, D.: Improving explainability and integrability of medical ai to promote health care professional acceptance and use: mixed systematic review. *Journal of Medical Internet Research* **27**, e73374 (2025)
- [43] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., et al.: Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* **106**, 102301 (2024)
- [44] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* **2**(1), 56–67 (2020)
- [45] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017)
- [46] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **267**, 1–38 (2019)
- [47] Mohseni, S., Block, J.E., Ragan, E.: Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*. pp. 22–31 (2021)

- [48] Pallathadka, H., Mustafa, M., Sanchez, D.T., Sajja, G.S., Gour, S., Naved, M.: Impact of machine learning on management, healthcare and agriculture. *Materials Today: Proceedings* **80**, 2803–2806 (2023)
- [49] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
- [50] Phillips-Wren, G.: Ai tools in decision making support systems: a review. *International Journal on Artificial Intelligence Tools* **21**(02), 1240005 (2012)
- [51] Raufi, B., Finnegan, C., Longo, L.: A comparative analysis of shap, lime, anchors, and dice for interpreting a dense neural network in credit card fraud detection. In: Longo, L., Lapuschkin, S., Seifert, C. (eds.) *Explainable Artificial Intelligence*. pp. 365–383. Springer Nature Switzerland, Cham (2024)
- [52] Rizzo, L., Longo, L.: A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning. In: *Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin, Dublin, Ireland, December 6-7th, 2018*. pp. 138–149 (2018)
- [53] Rossberg, N., Kleinberg, B., O’Sullivan, B., Longo, L., Visentin, A.: The feature understandability scale for human-centred explainable ai: Assessing tabular feature importance. *arXiv preprint arXiv:2510.07050* (2025)
- [54] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1**(5), 206–215 (2019)
- [55] Sahakyan, M., Aung, Z., Rahwan, T.: Explainable artificial intelligence for tabular data: A survey. *IEEE access* **9**, 135392–135422 (2021)
- [56] Scholbeck, C.A., Funk, H., Casalicchio, G.: Algorithm-agnostic feature attributions for clustering. In: *World Conference on Explainable Artificial Intelligence*. pp. 217–240. Springer (2023)
- [57] Shah, R., Pawar, A., Kumar, M.: Enhancing machine learning model using explainable ai. In: *International Conference on Data & Information Sciences*. pp. 287–297. Springer (2023)
- [58] Shapley, L.S., et al.: A value for n-person games (1953)
- [59] Sharifani, K., Amini, M.: Machine learning and deep learning: A review of methods and applications. *World Information Technology and Engineering Journal* **10**(07), 3897–3904 (2023)
- [60] Shuvra, M.K., Gony, M.N., Fatema, K.: Explainability vs. performance: Bridging the trade-off in deep learning models. *International Journal of Advanced Research in Computer Science & Technology (IJARCST)* **7**(5), 10931–10941 (2024)
- [61] Simuni, G.: Explainable ai in ml: The path to transparency and accountability. *International Journal of Recent Advances in* (2024)
- [62] Sovrano, F., Vitali, F.: An objective metric for explainable ai: how and why to estimate the degree of explainability. *Knowledge-Based Systems* **278**, 110866 (2023)
- [63] Štrumbelj, E., Kononenko, I., Šikonja, M.R.: Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* **68**(10), 886–904 (2009)

- [64] Veeramachaneni, K., O'Reilly, U.M., Taylor, C.: Towards feature engineering at scale for data from massive open online courses. arXiv preprint arXiv:1407.5238 (2014)
- [65] Vilone, G., Longo, L.: Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction* **3**(3), 615–661 (2021)
- [66] Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* **76**, 89–106 (2021)
- [67] Vilone, G., Longo, L.: Development of a human-centred psychometric test for the evaluation of explanations produced by xai methods. In: *World Conference on Explainable Artificial Intelligence*. pp. 205–232. Springer (2023)
- [68] Wang, Z., Huang, C., Yao, X.: A roadmap of explainable artificial intelligence: Explain to whom, when, what and how? *ACM Transactions on Autonomous and Adaptive Systems* **19**(4), 1–40 (2024)
- [69] Wiratsin, I.O., Ragkhitwetsagul, C.: Effectiveness of explainable artificial intelligence (xai) techniques for improving human trust in machine learning models: A systematic literature review. *IEEE Access* (2025)
- [70] Yang, F., Du, M., Hu, X.: Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint arXiv:1907.06831 (2019)
- [71] Zhao, H., Yu, S.: Cost-sensitive feature selection via the l_2 , l_1 -norm. *International Journal of Approximate Reasoning* **104**, 25–37 (2019)
- [72] Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **10**(5), 593 (2021)
- [73] Zhou, Q., Zhou, H., Li, T.: Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. *Knowledge-based systems* **95**, 1–11 (2016)
- [74] Zimmermann, R.S., Klein, T., Brendel, W.: Scale alone does not improve mechanistic interpretability in vision models. *Advances in Neural Information Processing Systems* **36**, 57876–57907 (2023)
- [75] Zini, J.E., Awad, M.: On the explainability of natural language processing deep models. *ACM Computing Surveys* **55**(5), 1–31 (2022)
- [76] Zytek, A., Arnaldo, I., Liu, D., Berti-Equille, L., Veeramachaneni, K.: The need for interpretable features: Motivation and taxonomy. *ACM SIGKDD Explorations Newsletter* **24**(1), 1–13 (2022)