Improving Explanations: Applying the Feature Understandability Scale for Cost-Sensitive Feature Selection

Nicola Rossberg[1,2][0009−0005−3883−5833], Bennett Kleinberg[4][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. With the growing pervasiveness of artificial intelligence, the ability to explain the inferences made by machine learning models has become increasingly important. Numerous techniques for model explainability have been proposed, with natural-language textual explanations among the most widely used approaches. When applied to tabular data, these explanations typically draw on input features to justify a given inference. Consequently, a user's ability to interpret the explanation depends on their understanding of the input features. To quantify this feature-level understanding, Rossberg et al. introduced the Feature Understandability Scale [49]. Building on that work, this proof-of-concept study collects understandability scores across two datasets, proposes a co-optimisation methodology of understandability and accuracy and presents the resulting explanations alongside the model accuracies. This work contributes to the body of knowledge on model interpretability by design. It is found that accuracy and understandability can be successfully co-optimised while maintaining high classification performances. The resulting explanations are considered more understandable at face value. Further research will aim to confirm these findings through user evaluation.

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

1 Introduction

In recent years, machine learning has become increasingly pervasive across a wide range of domains, including impactful areas such as medicine and agriculture, and its potential impact is undeniable [46, 44, 8, 54]. However, the risks associated with implementing black-box models, especially in critical fields such as medicine, must be acknowledged. The ability to audit machine learning systems is crucial to prevent the deployment of inequitable models, resulting from biased data or poor training. To promote the auditability of black- and grey-box machine learning models, a wide range of explainable AI (XAI) techniques has been proposed [62]. Through the implementation of XAI

techniques, model bias can be identified and mitigated, and the legal compliance and safety of machine learning models ensured [40].

One common form of explanation is textual, in which the model's inferences are expressed in natural language [59]. This is particularly advantageous for tabular data, where a model's output is explained in terms of its input features. For example, when applying for a loan, the input features may include 'income' and 'character'. If the loan were rejected, the textual explanation may state that "the loan was rejected due to 'income' being below 40,000 and 'character' being too low". Textual explanations are generally intuitive. However, their utility depends on users' ability to understand the features they include. In this example, the explanation is accessible only if the user understands the concepts of 'income' and 'character'. As such, it is in the system designer's best interest to prioritise understandable features in explanations to maximise their efficacy.

To quantify feature understandability, Rossberg et al. [49] proposed the *Feature Understandability Scale* (FUS), which measures feature understandability across various dimensions using user input. A final understandability score is assigned to each feature by averaging its ratings. To provide tailored explanations for the given user base, the scale should be completed by a subsample of end users to gauge the average understanding within the cohort. The understandability scores can then be used to co-optimise explanation quality and accuracy during machine learning model training. By favouring more understandable features during model design, the resulting explanations are expected to be more accessible and, consequently, more useful to the end user. To test this hypothesis and operationalise the scale, this proof-of-concept study implements the FUS to collect understandability ratings across two datasets, co-optimises the scores with accuracy, and assesses changes in the quality of the resulting explanations. To this end, three research questions are stipulated:

1. What are the trends in the distribution of FUS scores across two datasets?
2. How can accuracy and understandability be co-optimised in the machine learning workflow?
3. How does co-optimisation affect accuracy and explanation quality?

To answer these questions, a brief literature review of existing XAI methods and cost-sensitive feature selection is conducted in Section 2. Subsequently, user ratings of two datasets are solicited using the FUS. The datasets were selected based on their domain relevance, feature types and perceived understandability of features. Subsequently, the understandability scores are used to co-optimise accuracy and understandability. This process is detailed in Section 3. The final explanations are compared with those produced by a machine learning model optimised solely for accuracy, and the success of the proposed method is evaluated in Section 4.

2 Background

2.1 Explainable Artificial Intelligence

The pervasiveness of machine learning models across a range of domains has increased considerably over the past few years. However, despite the increase in processing power and the high accuracy of these models, the dangers associated with their black-box nature, including biases and shortcut learning, pose a significant barrier to their widespread implementation [10, 39, 56, 53]. One approach to address these concerns is to implement XAI techniques, which enable the audit of decision processes and have been shown to promote model transparency and trust [63]. To address the breadth of existing model architectures and provide a range of explanation types, a large number of XAI techniques have been proposed, including counterfactuals [36, 9, 29], feature attribution [50, 35], latent-space mechanistic interpretability [2, 70, 34, 1] and natural language methods [20, 71, 15]. A full review and taxonomy of XAI methods can be found at Vilone and Longo [60].

This study focuses on the use of textual natural-language explanations for tabular data. Textual explanations are advantageous because they are easily accessible and explain decision processes by referring to real variables. However, they also increase the potential for exaggerated user trust and confidence, as well as the anthropomorphisation of the algorithm, and therefore need to be situated correctly within the given context [21, 47, 65, 14]. Another drawback of textual explanations that has not previously been considered is that users' understanding of the explanation as a whole depends on their understanding of the input features it includes. Linking back to the previous example, a user's ability to understand why their loan was rejected depends on their understanding of the features 'income' and 'character'. To promote the use of understandable features, cost-sensitive feature selection approaches enable the co-optimisation of understandability and accuracy during machine learning training. Understandability scores are measured using the FUS [49]. The scale assesses feature-level understandability across two dimensions ('Feature-Outcome Relation' and 'Understanding and Measurement') and 8 or 9 items for numerical and categorical items, respectively. By deploying the scale to a sample of the explanation's target population, an average understandability score can be computed for each feature, and its understandability cost quantified.

2.2 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 [12, 38, 13]. 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 [69]. 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].

In the literature, three types of cost-sensitive feature selection methods are distinguished: filter methods, wrapper methods and embedded methods [67]. Filter methods

pre-select features before model training, and feature cost is used 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 often advantageous as they are computationally cheap and can be applied to high-dimensional datasets [6]. Wrapper methods use the performance of the machine learning model to evaluate feature selection. As such, feature subsets are evaluated by the same model, which is subsequently used for classification. This is advantageous, as the selected features are adapted to the learner and are therefore more accurate [6]. However, they are also more computationally expensive [32]. An example of this is the recursive feature elimination used in support vector machines [25, 67]. Both the filter and wrapper methods treat feature selection and classification as distinct steps. Embedded methods, on the other hand, combine feature selection and learner training within a single process. Feature cost is integrated into the machine learning model's loss function and optimised alongside performance during training [42, 24]. A summary of the three methods for cost-sensitive feature selection alongside their respective benefits and drawbacks is presented in Table 1.

Table 1: Summary of the benefits and drawbacks of the three cost-sensitive feature selection methods.

		+	−
Filter	Select features before training	Computationally cheap	Requires prior knowledge of costs
		Good for high-dimensional datasets	Ignores feature interaction
Wrapper	Select features through model performance	Features adapted to learner	More computationally expensive
		Accounts for feature interactions	Model-specific
Embedded	Integrate feature cost into loss function	Better performance	Not suitable for imbalanced data

The current work differs from the traditional literature on cost-sensitive feature selection in that feature costs are user-defined rather than error- or test-based. Where traditionally the feature cost is defined through its contribution to a costly misclassification or the cost of acquisition (e.g. the cost of a given medical test), feature cost here is operationalised as the inverse of understandability, stipulating that a costly feature is one that detracts from the understandability of the final explanations [38, 69, 13]. Additionally, the success of the final model is not assessed by misclassification rates (i.e., false positives and false negatives), but rather by the quality of the final explanation as evaluated by the target audience. Cost is operationalised as the inverse of the feature's understandability measured by the FUS. As such, a more costly feature is less understandable to the end-user; hence, understandable (and cheaper) features are prioritised during the cost-sensitive feature selection procedure. This study aims to minimise this cost by co-optimising it alongside accuracy. Unlike traditional literature, success is assessed by the final understandability of explanations and the accuracy of the co-optimised model.

2.3 Explanation Evaluation

The evaluation of explanation quality has been investigated in past literature, and several different approaches have been proposed [68, 7, 66, 43, 22]. Dai et al. [19] investigated how disparities in explanation quality may lead to misplaced trust and perpetuated bias and proposed a metric to evaluate disparities in explanations given a protected attribute. However, while this permits the evaluation of disparity within the model, it does not include a human-in-the-loop approach to evaluate the impact of the final explanations. Including human evaluation of explanation quality is helpful, as the perceived utility of explanations is user-dependent and the quality of an explanation should be evaluated by the group towards which it is targeted [51, 27].

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. 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 [18]. 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.

Such user-based evaluations should be collected consistently and reproducibly. In the field of psychometrics, this is possible through the development and validation of scales that capture user attitudes reliably [16]. Vilone and Longo [61] proposed such a 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 an explanation process. This allows the user to assess the extent to which the communication method makes the provided explanations accessible. While a range of explanation evaluation methods is available, none have been developed for analysing natural language explanations of tabular data. As this study aims to test the collection and co-optimisation of understandability scores, the quality of explanations will be evaluated through changes in understandability scores and qualitatively through example explanations.

2.4 Explainability-Accuracy Trade-off

The existence of an explainability-accuracy trade-off is contentious in the literature. On the one hand, it is argued that the implementation of more explainable grey-box models decreases accuracy due to their associated simplicity, and vice versa, complex black-box models lead to high performance but low interpretability [5, 31, 55, 33, 4]. The opposing argument states that this trade-off, while theoretically sound, is seldom observed in practice, in large part due to the lack of objective measurement of interpretability, and more transparent models should hence be favoured by default [11, 23, 17]. Additionally, user preferences and understanding of models may be domain-dependent, introducing a further consideration for model design [31]. In addition to the inherent interpretability

of a model, a range of post-hoc explainability methods such as GradCAM, SHapley Additive exPlanations (SHAP) and LIME have been proposed, which can provide explanations for model behaviour without impacting classification accuracy [52, 41, 48]. However, some of these methodologies may be less faithful to the system process, as they provide explanations through proxy models. As such, several important considerations need to be taken into account during model design to ensure suitable classification and explanation performance for the given task.

2.5 The Feature Understandability Scale

The FUS was proposed and validated by Rossberg et al. in [49]. The scale is a validated measurement instrument that quantifies the user's self-assessed understanding at the feature level. The FUS is divided into two subscales: one for categorical features and one for numerical features, and users complete the respective scale for each feature in a dataset. Users are defined as the target audience of the machine learning explanations based on the dataset. The scale is evaluated on a five-point Likert scale, and the mean rating is computed as the understandability score for each feature.

3 Methodology

The following details the data collection and co-optimisation of the understandability scores. This research aimed to assess the distribution of understandability scores across two domains and to implement a new methodology for their co-optimisation with accuracy. The overarching goal is to generate improved explanations on an ante-hoc basis. The section is structured according to the first three steps of the CRISP-DM guidelines [64]. These guidelines standardise the structure of data mining projects, promoting clear communication of the research methodology employed.

3.1 Business Understanding

This study aims to assess how understandability scores are distributed, how they can be integrated into machine learning training, and whether they enable the generation of more understandable explanations. From an end-user perspective, explanations are considered useful if they are both understandable and at the appropriate level [58]. As such, feature understanding should be adjusted to the specific target group of the explanations. For example, medical explanations targeted at patients need to differ from those targeted at practitioners. As such, the FUS should be completed by the target group. As a proof-of-concept, this work aims to test whether the implementation and co-optimisation of scores are feasible and beneficial, and therefore did not restrict participants to the target cohort. The success of the proposed methodology is evaluated by inspecting the feature understandability scores of both the traditional and co-optimised explanations, as well as the accuracy of the respective models. This aims to show both the degree to which understandability improves and the overlap between important and understandable features. The project will be structured in three key phases. First, two datasets from different domains are selected, and the understandability scores of

their features are acquired. Second, accuracy and understandability are co-optimised, and both traditional and improved explanations are generated. Finally, the generated co-optimised explanations are qualitatively evaluated and compared to the traditional explanations. The exact workflow of the paper is showcased in Figure 1. The figure shows the collection of understandability ratings in the first step of the study. During the co-optimisation procedure, the ten best features are selected using traditional feature selection; of those, the five most understandable are retained and classified using machine learning models. In the traditional feature selection comparison, the five best features are selected directly from the dataset and used to classify. Local explanations for both models are generated, and the accuracy and understandability scores are presented.

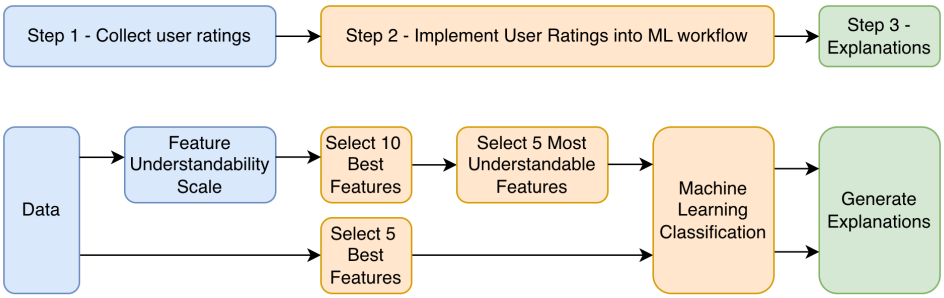

Fig. 1: Workflow of this study.

3.2 Data Understanding

The FUS is applied to two public datasets to ensure that any changes in explanation quality are not domain-specific. The first dataset is the 'Telco Customer Churn' dataset, which contains data that can be used to predict whether a customer will switch service providers given their family and usage history. The dataset is publicly available [5]. This dataset will be referred to as 'Dataset 1' 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 [6] and will be referred to as 'Dataset 2' going forward. Before collecting the understandability scores, 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.

Understandability scores were collected using a survey on the platform Qualtrics[7]. Data collection was approved by the University College Cork Social Research Ethics

[5] `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`
[6] `https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking`
[7] `https://www.qualtrics.com/`

Committee under Log Number 2025-231 on the 1st of October 2025. Participants were recruited via convenience and snowball sampling, and no personal data were collected. Snowball sampling was used by asking participants to pass the survey to their networks, thereby exponentially increasing the reach of the data collection. Participants had to be over 18 and have a self-assessed English level of B1 or higher. All participants had to provide informed consent before commencing the study. Participants were assigned to one of the two datasets and asked to rank half of the features using the FUS. Each participant ranked 10 (Dataset 1) or 12 (Dataset 2) features using the scale. 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 equal number of ratings. Participant responses were cleaned, and fast responders were removed from the dataset. Additionally, all responders with a self-reported English Level below B1 were removed. The understandability scores were then reverse-coded to yield cost values ranging from 1 (lowest cost, most understandable) to 5 (highest cost, least understandable). Where previously more understandable features were given higher ratings, the reverse coding reverses the scale, assigning a lower cost to more understandable features and vice versa. This is done to align the methodology with the existing literature on cost-sensitive feature selection. As such, more costly (and less understandable) features are considered more expensive to the acquisition process and cheaper (more understandable) features are favoured. Finally, the mean understandability cost was computed for each feature.

3.3 Modelling

After data collection, the understandability cost scores were integrated into the machine learning pipeline using a filter method, enabling the creation of improved explanations on an ante-hoc basis. A filter method was selected for the co-optimisation procedure to ensure transparent feature selection, permit replication and minimise confounding effects that detract from the goal of the current research. First, a preliminary model selection was conducted using stochastic resampling to test the implementation and evaluate the performance of a Decision Tree, a Random Forest, and a Support Vector Machine (SVM) on the training set, with the models retained based on average classification performance. For each model, Select K Best with chi-square, f-classif, and mutual-info-classif, as well as Recursive Feature Elimination, were tested and retained based on the highest cross-validation accuracy in the training set. Chi-Square computes the exponential chi-square kernel, F-classif the ANOVA F-value, and mutual-info-classif computes the mutual information based on k-nearest neighbours estimation between each feature and the target variable. Based on these respective information scores, Select K Best ranks all features and retains the K best features, where K is user-specified. Recursive Feature Elimination is a more complex procedure in which a model (here, a Decision Tree) is trained, and the assigned feature importances are used to iteratively remove features and recompute them. All feature selection methods were implemented through Scikit-learn [45]. After selecting the machine learning and feature selection methods, accuracy and understandability are co-optimised to minimise the potential trade-off between explainability and accuracy. In the literature, it is argued that this trade-off may result from simpler models being more transparent but yielding lower accuracy levels due to

their reduced complexity [33]. The current method is model-agnostic for processing tabular data and is therefore not expected to directly impact performance. However, the trade-off may result from non-understandable features being key to discriminating between classes, while understandable features may contain less information. As such, the introduction of understandability-based feature selection may lead to the exclusion of important features and the reduction of model performance. To counteract this, the proposed method aims to co-optimise accuracy and understandability, minimising predictive power loss while promoting the use of understandable features.

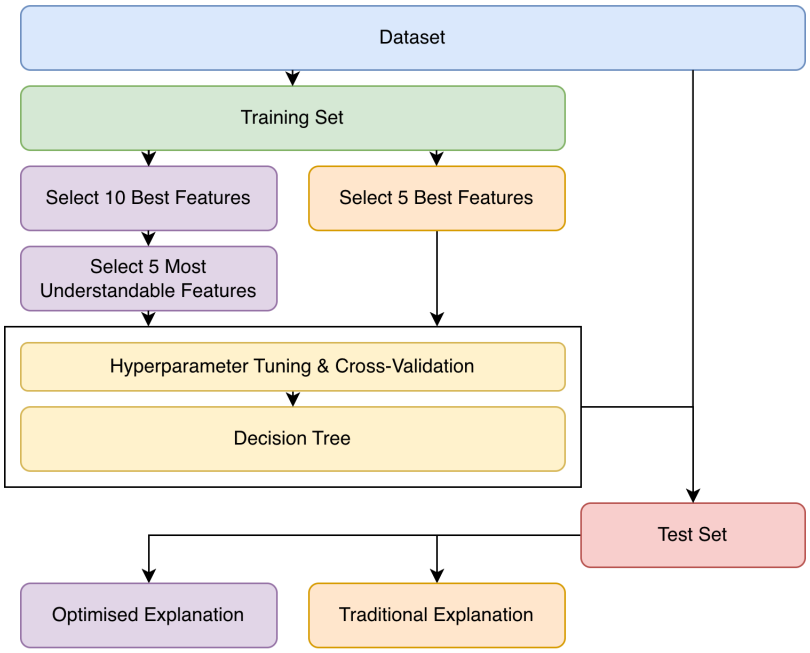

Fig. 2: Workflow of the Cost-Sensitive feature selection

The workflow of the implemented co-optimisation is depicted in Figure 2. The datasets were split into training and test sets with a stratified 70-30 split. The training set was further split into training and validation with a stratified 70-30 split for the model and feature selection algorithm selection procedure. Numerical features were standardised, and one-hot and label encodings were implemented for multiclass and binary categorical features, respectively. In the first step of the co-optimisation, the ten most important features were selected through traditional feature selection. Four feature selection algorithms (Select K Best with the chi-square, f-classif, and mutual-info-classif kernels, and Recursive Feature Elimination) were tested on the training data for each dataset and model, and the best-performing algorithm was retained. Of the ten most important features selected, the five most understandable were retained for

classification. Three classification models (SVM, Random Forest, and Decision Tree) were tested on the training set, and the two best-performing models were implemented on the validation data. The training and validation data were then merged, and the 5-feature dataset was classified, using a grid search with 5-fold cross-validation for model tuning. Both feature selection and classification were implemented through the Scikit-learn library [45]. The best model was retrained on the entire training data and evaluated on the reserved testing set. As a comparator to the co-optimisation procedure, traditional feature selection was implemented. Here, the five most important features were selected from the training dataset using the same feature selection algorithm. A machine learning model was then trained using the same grid search and 5-fold cross-validation procedure and evaluated on the reserved test data. Due to the range of feature selection and classification algorithms tested, evaluation was conducted only using a stratified train-test split.

After model tuning and evaluation, SHAP was implemented to extract feature importances and local explanations for the chosen models [41]. While some models (e.g., Decision Tree) are inherently transparent and provide more faithful ways to extract feature importances, SHAP was chosen to enable cross-model comparison. The selected SHAP model was 'Tree Explainer'. SHAP computes feature importance by averaging the change in model output when each feature is introduced one at a time, conditioning on the feature in question. The Tree Explainer is specifically tailored to tree- and ensemble-tree models, making it suitable for the current study. Feature importances were computed for a selected range of test set instances and transformed into explanations. A heuristic was created to generate examples of the final explanations. Several considerations were made during the design process, including the length of explanations, the provision of comparators for individual values, and the direction in which the individual features influenced the decision. The final explanation was structured to first present the predicted outcome, then introduce the features in favour of it, followed by those that detract from the decision. For each feature, the instance's value was provided alongside the mean (for numerical features) to contextualise the instance's position. The features in the explanations are ordered by their relative importance to the final prediction. The full code, including feature selection, model training and explanation generation, can be found at `github.com/ncrossberg/User-Study`.

4 Results and Discussion

4.1 Feature Rating Collection

Feature ratings were collected from the 15th of October to the 1st of November 2025. A total of 163 responses were initially collected. After removing very fast respondents (less than 3 minutes) and incomplete surveys, 54 respondents remained. As each respondent rated one quarter of the features, the average number of ratings per feature was 13.53. The distributions of ratings across the two datasets, along with their standard deviations, are shown in Figures 3 and 4. The distributions of ratings were rather small: Dataset 1 ranged from a minimum cost of 1.70 (Contract) to 2.88 (Partner), while Dataset 2 ranged from a minimum cost of 1.88 (Age) to a maximum of 3.38 (hearing (left)). On the other hand, the scores' standard deviation is relatively large, indicating considerable variation within the response distribution for each feature.

Several reasons for the limited range of the results can be postulated. However, these are only theoretical assertions, and an item response or diffusion model is necessary to confirm them. One possibility is that this is an instance of 'fence sitting', a phenomenon in which respondents choose neutral answers despite having a stronger opinion. Potential reasons for fence-sitting include social desirability (claiming to understand a feature even if they don't), apathy (disinterest in the questionnaire), indecisiveness, or lack of information [37]. While fence-sitting cannot be avoided entirely, future research may wish to include further clarifying information, as well as attention checks and financial incentives to combat this behaviour. An alternative explanation for this phenomenon is that the range of feature difficulty was, in fact, limited, and the ratings are representative of a dataset that, on average, is understandable. The high variance within each feature's ratings may be attributed to three potential causes. First, as a pilot study, participants were not limited to the population of interest, leading to a more diverse sample. This may cause considerable discrepancies in understanding, increasing the variance. In addition, the small sample size of this study may also increase variance. As the number of participants increases, the impact of random fluctuations decreases, and the variance of results tends to decrease. However, this is not guaranteed, as the variance may reflect a genuine discrepancy in understanding within the population of interest. A third possibility is that participants may have misunderstood the scale labels, hence introducing error to the measurements. As such, future research should clearly define the scale's endpoints and explore whether the high variance persists in larger samples.

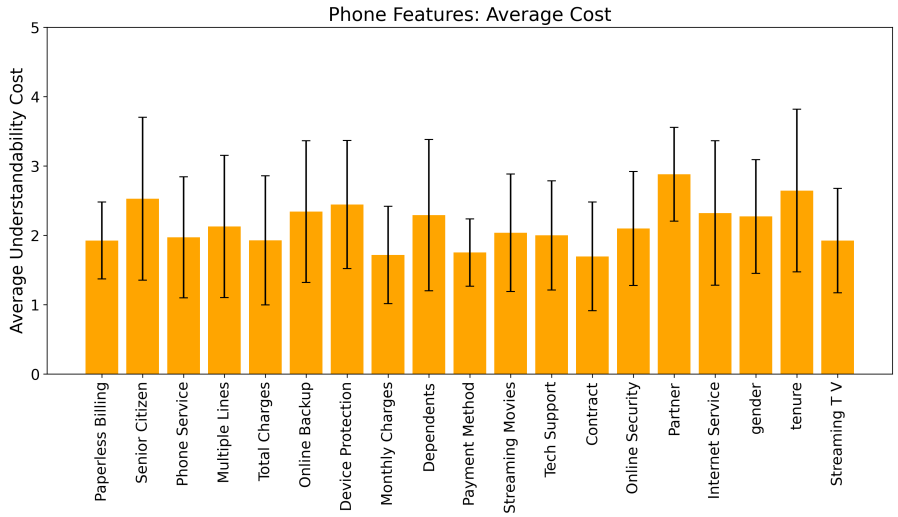

Fig. 3: Average feature cost and standard deviation for the phone company customer churn dataset. Note that cost is computed as the inverse of understandability to permit co-optimisation. A lower score indicates a more understandable feature.

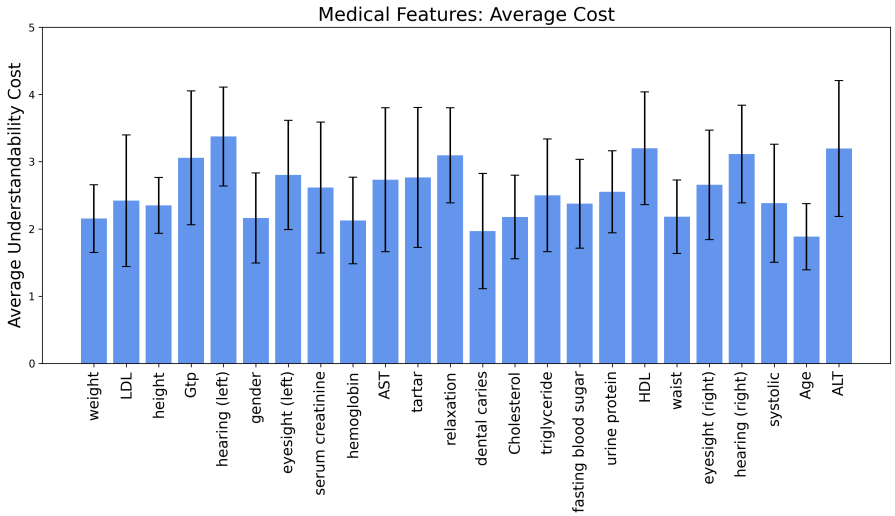

Fig. 4: Average feature cost and standard deviation for the medical dataset. Note that cost is computed as the inverse of understandability to permit co-optimisation. A lower score indicates a more understandable feature.

On average, Dataset 2 was found to be more difficult, which is in line with expectations, as respondents are likely to be less familiar with specialised medical terms. However, it is important to note that direct comparisons between the datasets are not recommended. Participants are likely to use other features within the same dataset as internal reference points when providing ratings. As a result, their interpretation of the rating scale and their assessment of feature difficulty may differ across datasets. Because each participant was exposed to only one dataset, these relative judgments are not anchored to a common baseline, preventing a meaningful direct comparison of understandability scores across datasets. Additionally, presenting a random subset of features in each dataset may contribute to the high variance. If, by chance, a participant receives only difficult features, the more understandable features would be scored as comparatively costly due to the misaligned reference. Surprisingly, the most costly feature in Dataset 2 was 'hearing (left)', which was considered to be reasonably understandable at face value. However, the high cost likely stems from the measurement scale, which respondents may be unfamiliar with. A similar phenomenon is observed in Dataset 1, where 'Partner' is the most costly feature, despite being understandable at face value. One conclusion that may be drawn from this is that open, easily accessible documentation of rating scales is vital for understanding features. Future research should provide a key for the feature values during rating collection if the measurement scale is ambiguous. In addition, larger samples should be collected to improve generalisation and decrease variance by minimising the impact of outliers. The exact number of samples necessary will depend on the domain and aim of the study [57].

4.2 Co-Optimisation

The results of the co-optimisation alongside the traditional feature selection are presented in Table 2. After testing several feature selection and classification methods, Random Forest and Decision Tree were retained as classifiers based on their validation accuracy. The features in Dataset 1 were selected using the Select K best algorithm with an f-classif kernel for the Decision Tree, and Recursive Feature Elimination with a Decision Tree as the model for the Random Forest. Those in Dataset 2 were selected using Recursive Feature Elimination with a Decision Tree as the model for both the Decision Tree and the Random Forest.

The classifiers in both datasets achieve good balanced accuracies (73.08 - 75.26% for Dataset 1 and 75.49 - 76.05% for Dataset 2) with minimal overfitting, implying that the models are correctly specified for the data and classification task. In three out of four cases, test accuracy is slightly lower in the co-optimised compared to the traditional model. Only the Random Forest in Dataset 1 performs slightly better under co-optimisation conditions, which may be attributed to sample variations or the co-variation of understandability and feature importance. The slight drop in accuracy observed across the remaining three cases is in line with the expected accuracy-explainability trade-off, where important features may score low on understandability. Importantly, the total and mean understandability costs decrease in the co-optimised model, highlighting that the proposed method successfully selects more understandable features while retaining relatively high accuracy levels. The decrease in average cost is relatively small (0.31 (Decision Tree) and 0.3 (Random Forest) in Dataset 1 and 0.36 (Both Models) in Dataset 2. However, relative to the overall range in cost, it constitutes 26.2% (Decision Tree) and 25.4% (Random Forest) of the total variation in cost for Dataset 1 and 24% (both models) for Dataset 2. As such, understandability increases considerably compared to the relative drop in accuracy, emphasising that it is possible to develop models that remain accurate while prioritising understandable features.

When comparing the two models, the Random Forest performs better in Dataset 1, demonstrating both higher accuracy values and lower feature cost. However, the model also requires significantly longer training and prediction times, and the accuracy-time-understandability trade-off should be considered on a case-by-case basis. In Dataset 2, the Decision Tree performs better in terms of accuracy, with both models selecting the same features and reporting equal understandability scores, while the Random Forest has considerably larger training and prediction times. As such, the FUS permits the consideration of not only run-time and accuracy during model selection but also the integration of understandability into this procedure.

The features selected by each model across the two datasets are presented in Table 3. Two features (highlighted in bold) are selected by both the traditional and the co-optimised Decision Tree in both datasets, implying that they are both important and understandable. While it is unlikely that this pattern will generalise across all domains and datasets, the observed correlation between feature importance and understandability is remarkable. This pattern presents even more strongly in the Random Forest, where three of the same features are selected in Dataset 1. Interestingly, this selection of similar features also co-occurs with improved understandability, further supporting the correlation between informativeness and understandability in the chosen datasets.

Table 2: Balanced Accuracy (in %) and Training Times (in seconds) for the Decision Tree and Random Forest models for Dataset 1 (customer churn) and Dataset 2 (smoking prediction)

Decision Tree	Dataset 1		Dataset 2	
	Traditional	Co-Optimisation	Traditional	Co-Optimisation
Train Accuracy	76.94	74.74	76.64	76.11
CrossVal Accuracy	76.81	74.74	76.27	75.99
Test Accuracy	74.56	73.08	76.05	75.57
Total Cost	10.76	9.18	12.8	11.02
Mean Cost	2.15	1.84	2.56	2.20
Training Time	0.0052	0.025	0.021	0.020
Prediction Time	0.0017	0.014	0.0033	0.0032
Random Forest	Dataset 1		Dataset 2	
	Traditional	Co-Optimisation	Traditional	Co-Optimisation
Train Accuracy	77.17	76.53	76.08	75.90
CrossVal Accuracy	77.06	75.92	75.85	75.85
Test Accuracy	75.05	75.26	75.7	75.49
Total Cost	10.31	8.79	12.80	11.02
Mean Cost	2.06	1.76	2.56	2.20
Training Time	0.81	0.79	3.14	2.67
Prediction Time	0.034	0.035	0.20	0.20

Within the domains, the features that are both important and understandable make sense intuitively. In Dataset 1, the Decision Trees select a *month-to-month contract* with *no tech support*. At face value, these features are both understandable and expected to be strongly correlated with customer churn, given the short-term, low-support nature of such contracts. The Random Forest also selects the *month-to-month contract* alongside the *monthly* and *total charges* in both the traditional and co-optimised settings. As the only two numerical features in the dataset, *monthly charges* and *total charges* are understandable at face value and likely informative, given their strong correlation with customer loyalty (the longer a customer has stayed with a firm, the higher the total charges). Similarly, in Dataset 2, the features selected by the Decision Tree and Random Forest are *gender* and *age*, both of which are understandable and expected to strongly correlate with whether a patient smokes. Interestingly, the two features selected by both models in Dataset 1 show considerable differences in importance between the traditional and co-optimised models. In Dataset 2, on the other hand, *gender* and *age* are the two most important features in both models, further emphasising their strong predictive and explanatory relevance.

The final generated explanations are shown in Table 4. At face value, the features selected in the co-optimised models appear more understandable than those in the traditional model, with the included terminology being more commonly accessible. One notable aspect of the explanations in Dataset 1 for both models is the use of one-hot encoding for categorical features. Several variables in this dataset (e.g., Contract and

Table 3: Feature importances for both the traditional and co-optimised feature selection methodologies for Dataset 1 (customer churn) and Dataset 2 (smoking prediction). Features are ordered from most to least important.

Decision Tree			
Dataset 1		Dataset 2	
Traditional	Co-Optimisation	Traditional	Co-Optimisation
Tenure,	**Contract: Month-to-month**,	**Gender**,	**Age**,
Internetservice: Fiber Optic,	Contract: Two year,	**Age**,	**Gender**,
Onlinesecurity: No,	Payment Method: Electronic check,	Triglyceride,	Waist,
Techsupport: No,	**Techsupport: No**,	ALT,	Fasting Blood Sugar,
Contract: Month-to-month	Streaming Movies: No internet service	GTP	LDL
Random Forest			
Traditional	Co-Optimisation	Traditional	Co-Optimisation
Tenure,	**Contract Month-to-month**,	**Gender**,	**Age**,
Monthly Charges,	Contract One year,	**Age**,	**Gender**,
Total Charges,	**Monthly Charges**,	Triglyceride,	Waist,
Internet Service Fiber Optic,	Payment Method: Electronic check,	ALT,	Fasting Blood Sugar,
Contract Month-to-month	**Total Charges**	GTP	LDL

Online Security) had multiple categories and were therefore one-hot encoded before analysis, with each encoded category retaining the same cost as the original feature. While this representation is suitable for model training, it complicates the interpretation of the resulting explanations. Specifically, the explanation may emphasise the absence of a particular category rather than the participant's actual category membership. For example, instead of stating that a participant has a month-to-month contract, the explanation may highlight that the participant does not have a one-year contract. This framing may be unintuitive for users and may require an implicit understanding of how one-hot encoding represents categorical data. As a result, overall interpretability depends not only on the user's understanding of the original feature but also on understanding its individual categories and the relationships between them. Furthermore, presenting multiple binary indicators for a single categorical variable can fragment the information and make explanations less accessible, particularly for lay users who may not be familiar with one-hot encoding. Future research may wish to mitigate these issues by weighting the scale items that assess understanding of individual categories more heavily during FUS evaluation. Alternatively, aggregating one-hot-encoded categories before explanation creation could be explored. While this would likely make explanations more accessible, it would also decrease faithfulness, and the more suitable choice will likely be domain and context-dependent.

In Dataset 2, the co-optimised features are somewhat more understandable, as only 2 are specialised medical terms compared to three in the traditional explanation. However, it is unclear whether the quantity or presence of non-understandable features has a greater impact on user understanding, and future research may wish to clarify this. An interesting aspect of the explanations in Dataset 2 is the use of acronyms (e.g. ALT, GTP, LDL). While acronyms may at times be more commonly known than the words they represent, they may also reduce the understandability of explanations, depending

on the domain and user base. As such, it may be of interest to researchers to present abbreviations alongside the full names of variables during the understandability score collection to prevent any related ambiguity.

However, all of these hypotheses need to be verified through specific user feedback to assess the real impact of the explanation structure and presentation. In conjunction with quantitative evaluation via a scale, qualitative interviews may be useful for identifying other areas for improvement in the structure and composition of the explanations.

Table 4: Example explanations of the co-optimised and traditional feature selection for both datasets. These explanations were created in the last step of the co-optimisation procedure through the feature importances extracted from SHAP and the designed explanation heuristic.

	Decision Tree
	Co-Optimised
Dataset 1	The customer is predicted to stay with the company because their contract category is Two year and their streaming movies category is No internet service. The fact that their contract category is not Month-to-month and their payment method category is not Electronic check and their techsupport category is not No detracted from this prediction.
Dataset 2	The individual is predicted to be a smoker because their gender is M and their age of 40.00 is less than average and their waist of 98.00 is more than average. The fact that their fasting blood sugar of 106.00 is more than average and their ldl of 106.00 is less than average detracted from this prediction.
	Traditional
Dataset 1	The customer is predicted to stay with the company because their tenure of 6487.79 is more than average and their contract category is Two-year. The fact that their online security category is not No and their techsupport category is not No and their contract category is not Month-to-month detracted from this prediction.
Dataset 2	The individual is predicted to be a smoker because their gender is M, their age of 40.00 is less than average, their gtp of 127.00 is more than average, their alt of 24.00 is less than average. The fact that their triglyceride of 89.00 is less than average detracted from this prediction.
	Random Forest
	Co-Optimised
Dataset 1	The customer is predicted to stay with the company because their monthly charges of 165.83 are less than average. The fact that their contract category is not Month-to-month, their contract category is not One year, their payment method category is not Electronic Check, their Total Charges of 370.50 are less than average detracted from this prediction.
Dataset 2	The individual is predicted to be a smoker because their gender is M, their waist of 98.00 is more than average, their age of 40.00 is less than average, and their fasting blood sugar of 106.00 is more than average. The fact that their ldl of 106.00 is less than average detracted from this prediction.
	Traditional
Dataset 1	The customer is predicted to stay with the company because their tenure of 2093.76 is less than average and their monthly charges of 165.83 are less than average, and their total charges of 370.50 are less than average. The fact that their internet service category is not Fiber optic, and their contract category is not Month-to-month detracted from this prediction.
Dataset 2	The individual is predicted to be a smoker because their gender is M, their gtp of 127.00 is more than average, their age of 40.00 is less than average, and their alt of 24.00 is less than average. The fact that their triglyceride of 89.00 is less than average detracted from this prediction.

5 Limitations and Future Work

Despite the best efforts of the authors, several limitations of this study can be conceived, and possible avenues for future work are proposed on this basis. The categories of limitations can be divided into data collection, co-optimisation, explanation evaluation, and possible improvements and extensions for each venue, which are detailed below.

5.1 Data Collection

This was the first study to implement the FUS to collect understandability scores and analyse their attributes. Based on the distribution and variation of the scores, several conclusions can be drawn. First, the limitations of the small dataset in a proof-of-concept study need to be acknowledged. As a result, the variance across feature ratings is relatively large, and the representativeness of the scores is uncertain. As such, future research should implement the methodology established in this study on larger FUS datasets. As this was the first instance of FUS usage, several recommendations for future understandability score collection can be made based on observed response behaviour. The first encountered problem was fence-sitting, where users overwhelmingly selected neutral scores, resulting in relatively narrow ranges in understandability. To combat this, future research can include clarifying information, attention checks, and financial incentives to improve the quality of collected responses. Additionally, some features that appeared simple at face value proved relatively difficult for users, likely due to confusion about their measurement scales. As such, future research may wish to provide clarifying information about measurements when scales are deemed ambiguous. Finally, data collection through snowball sampling may introduce sampling biases depending on the domain and network of the researcher, and future research should aim to use random sampling practices.

The current research was limited to two datasets in relatively accessible domains. The collected explainability scores were distributed within a small range centred in the middle of the scale, with no features being considered excessively easy or difficult. While other domains (e.g. medicine or finance) may introduce less understandable features, the ratings are collected in the context of the given domain. As such, users will complete ratings relative to other items in the given dataset, and ratings are expected to remain somewhat centred within a given domain.

5.2 Co-Optimisation

The current study took a filter approach to the co-optimisation of understandability and accuracy. While this permitted the maintenance of high accuracy levels while promoting 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. An example is that the co-optimised explanations for Dataset 2 still contain two specialised medical terms, and it may be of interest to assess whether the presence or quantity of difficult features is more impactful for explanation understanding.

In addition to implementing other cost-sensitive feature selection methods, future research may wish to consider the implementation of more complex machine learning models. This study tested Random Forests, Decision Trees and SVMs for classification, but it may be of interest to assess whether the accuracy-understandability trade-off can be minimised with more complex models such as Gradient Boosting or small neural networks designed for tabular data, such as TabNet [3]. While these networks carry some implicit challenges, such as decreased transparency, the integration of understandability and usage of ad-hoc explainability methods should combat these problems.

Additionally, this research assessed the feature selection and classification performance of all models using only a stratified train-test split. Future research may wish to investigate the stability of selected features across random seats and to implement cross-validation or bootstrapping to assess their effects on both the selected features and changes in understandability scores. This would also permit the computation of the statistical significance of the changes.

Another possible extension of the current research is the introduction of a balancing coefficient adjusting the prioritisation of understandability and accuracy. While the proposed filter-based method does not allow for this, an embedded co-optimisation may enable it by introducing an understandability coefficient into the loss function, which adjusts the impact of understandability scores during model optimisation. This would allow prioritising understandability and accuracy scores by domain and target audience, thereby enabling a more flexible integration of understandability into the machine learning training process.

5.3 Explanation Evaluation

Based on the generated explanations, it is notable that features that require one-hot encoding may be more difficult to understand, depending on the category labels and the user's understanding of the encoding logic. As such, it may be of interest to weigh measurement-scale and category-understanding questions more heavily in the computation of the understandability score for features that require one-hot encoding. This would ensure that categories included in the final explanation do not decrease understandability. Alternatively, it may be worthwhile to explore the concatenation of one-hot encoded variables into the parent feature before the creation of explanations. However, this would decrease the explanation's faithfulness, which is a trade-off that may not be desirable depending on the work's context and domain. However, the most important next step is to collect user feedback to evaluate the change in explanation quality resulting from the co-optimisation procedure. While explanations were assessed through changes in understandability cost and at face value in the current study, user feedback is necessary to ensure that the quality improvements are both present and effective in the end-user base. Ideally, explanations should be evaluated using an assessment scale and compared with traditional explanations to determine which users prefer. The authors are currently implementing a validation study to evaluate changes in quality. To this end, future studies would also be of interest to develop and validate a scale to measure the understandability and quality of textual explanations. While similar scales assessing various dimensions of explanation quality exist, none have been found to specifically evaluate the quality of textual explanations.

6 Conclusion

This proof-of-concept study aimed to implement the FUS on two datasets, analyse the score distribution, propose a co-optimisation technique and evaluate the resulting explanations. The FUS scores showed a limited range of feature understandability across both datasets, and reasons as well as possible solutions for this were proposed. The accuracy in co-optimised models decreased somewhat compared to the traditional setting. However, the co-optimised models also successfully promoted the use of understandable features, decreasing the mean and total understandability cost and generating more understandable explanations at face value. These findings highlight that improved explanations can be generated at low accuracy cost through the promotion of understandable features. However, further research collecting user feedback on the explanations' quality is necessary to assess whether there is a measurable change in quality.

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.

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. vol. 35, pp. 6679–6687 (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., Véras, 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] 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)

[12] 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)

[13] 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)

[14] Brdnik, S., Colakovic, I., Karakatič, S.: Non-experts' trust in xai is unreasonably high. In: World Conference on Explainable Artificial Intelligence. pp. 184–197. Springer (2025)
[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] 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)
[19] Dai, J., Upadhyay, S., Aivodji, U., Bach, S.H., Lakkaraju, H.: Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 203–214 (2022)
[20] 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)
[21] DeVrio, A., Cheng, M., Egede, L., Olteanu, A., Blodgett, S.L.: A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–18 (2025)
[22] 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)
[23] 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)
[24] Gao, W., Hu, L., Zhang, P.: Class-specific mutual information variation for feature selection. Pattern Recognition **79**, 328–339 (2018)
[25] Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1), 389–422 (2002)
[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] Hilton, D.J.: Conversational processes and causal explanation. Psychological Bulletin **107**(1), 65 (1990)
[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] Kästner, L., Crook, B.: Explaining ai through mechanistic interpretability. European journal for philosophy of science **14**(4), 52 (2024)
[35] 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)
[36] 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)
[37] Kulas, J.T., Stachowski, A.A.: Middle category endorsement in odd-numbered likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. Journal of Research in Personality **43**(3), 489–493 (2009)
[38] 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)
[39] 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)
[40] 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)
[41] 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), 2522–5839 (2020)
[42] Ma, X.A., Xu, H., Liu, Y., Zhang, J.Z.: Class-specific feature selection using fuzzy information-theoretic metrics. Engineering Applications of Artificial Intelligence **136**, 109035 (2024)
[43] 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)
[44] 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)

[45] 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)

[46] Phillips-Wren, G.: Ai tools in decision making support systems: a review. International Journal on Artificial Intelligence Tools **21**(02), 1240005 (2012)

[47] Placani, A.: Anthropomorphism in ai: hype and fallacy. AI and Ethics **4**(3), 691–698 (2024)

[48] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

[49] 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)

[50] 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)

[51] Schuff, H., Adel, H., Qi, P., Vu, N.T.: Challenges in explanation quality evaluation. arXiv preprint arXiv:2210.07126 (2022)

[52] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

[53] 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)

[54] 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)

[55] 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)

[56] Simuni, G.: Explainable ai in ml: The path to transparency and accountability. International Journal of Recent Advances in (2024)

[57] Sinha, A., Nayem, H., Kibria, B.G.: Sample size requirements for the central limit theorem for skewed distributions: A simulation study. Journal of Statistical Modeling and Analytics (JOSMA) **7**(2) (2025)

[58] Verdejo, V.M., Quesada, D.: Levels of explanation vindicated. Review of Philosophy and Psychology **2**(1), 77–88 (2011)

[59] Vilone, G., Longo, L.: Classification of explainable artificial intelligence methods through their output formats. Machine Learning and Knowledge Extraction **3**(3), 615–661 (2021)

[60] Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion **76**, 89–106 (2021)

[61] 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)
[62] 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)
[63] 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)
[64] Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. vol. 1, pp. 29–39. Manchester (2000)
[65] Xuan, Y., Small, E., Sokol, K., Hettiachchi, D., Sanderson, M.: Comprehension is a double-edged sword: Over-interpreting unspecified information in intelligible machine learning explanations. International Journal of Human-Computer Studies **193**, 103376 (2025)
[66] Yang, F., Du, M., Hu, X.: Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint arXiv:1907.06831 (2019)
[67] Zhao, H., Yu, S.: Cost-sensitive feature selection via the 2, 1-norm. International Journal of Approximate Reasoning **104**, 25–37 (2019)
[68] 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)
[69] 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)
[70] 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)
[71] Zini, J.E., Awad, M.: On the explainability of natural language processing deep models. ACM Computing Surveys **55**(5), 1–31 (2022)