THE FIX BENCHMARK: EXTRACTING FEATURES INTERPRETABLE TO EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature-based methods are commonly used to explain model predictions, but these methods often implicitly assume that interpretable features are readily available. However, this is often not the case for high-dimensional data, and it can be hard even for domain experts to mathematically specify which features are important. Can we instead automatically extract collections or groups of features that are aligned with expert knowledge? To address this gap, we present FIX (Features Interpretable to eXperts), a benchmark for measuring how well a collection of features aligns with expert knowledge. In collaboration with domain experts, we propose FIXScore, a unified expert alignment measure applicable to diverse real-world settings across cosmology, psychology, and medicine domains in vision, language and time series data modalities. With FIXScore, we find that popular feature-based explanation methods have poor alignment with expert-specified knowledge, highlighting the need for new methods that can better identify features interpretable to experts.

1 INTRODUCTION

Machine learning is increasingly used in domains like healthcare (Tjoa & Guan, 2019), law (Atkinson et al., 2020), governance (Meijer & Wessels, 2019), science (de la Torre-López et al., 2023), education (Holstein et al., 2018) and finance (Modarres et al., 2018). However, modern models are often black-box, which makes it hard for practitioners to understand their decision-making and safely use model outputs (Rai, 2019). For example, surgeons are concerned that blind trust in model predictions will lead to poorer patient outcomes (Hameed et al., 2023); in law, there are known instances of wrongful incarcerations due to over-reliance on faulty model predictions (Zeng et al., 2016; Wexler, 2017). Although such models have promising applications, their opaque nature is a liability in domains where transparency is crucial (Jacovi et al., 2021; Hong et al., 2020).

To address the pertinent need for transparency and explainability of their decision-making, the interpretability of machine learning models has emerged as a central focus of recent research (Arrieta et al., 2019; Saeed & Omlin, 2023; Räuker et al., 2023). A popular and well-studied class of interpretability methods is known as *feature attributions* (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017). Given a model and an input, a feature attribution method assigns scores to input features that reflect their respective importance toward the model's prediction. A key limitation, however, is that the attribution scores are only as interpretable as the underlying features themselves (Zytek et al., 2022).

Feature-based explanation methods commonly assume that the given features are already interpretable to the user, but this typically only holds for low-dimensional data. With high-dimensional data like images and text documents, where the readily available features are individual pixels or tokens, feature attributions are often difficult to interpret (Nauta et al., 2023). The main problem is that features at the individual pixel or token level are often too granular and thus lack clear semantic meaning in relation to the entire input. Moreover, the important features are also domain-dependent, which means that different attributions are needed for different users. These factors limit the usefulness of popular feature attribution methods on high-dimensional data.

Instead of individual features, people understand high dimensional data in terms of semantic collections of low level features, such as regions in an image or phrases in a document. Moreover, for a feature to be useful, it should align with the intuition of *domain experts* in the field. To this end, an interpretable feature for high-dimensional data should have the following properties. First,

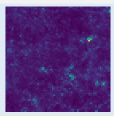
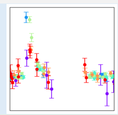
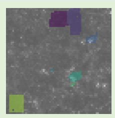
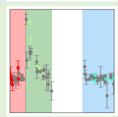
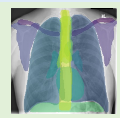
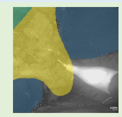
	Implicit Expert Features				Explicit Expert Features	
	Cosmology		Psychology		Medicine	
Dataset	Mass Maps	Supernova	Multilingual Politeness	Emotion	Chest X-Ray	Cholecystectomy
Input (x)	mass map image	simulated astronomical time-series data	conversation snippet	Reddit comment	chest X-ray image	video surgery image
Output (y)	energy density Ω_m, matter fluctuation σ_8	astronomical sources (e.g. supernova)	politeness level	emotion	pathology	safe/unsafe zone
# Examples	110,000	7,848	22,800	58,000	28,868	1,015
Expert Features	voids, clusters	linear consistent wavelengths	lexical categories	Russell's circumplex model	anatomical structures	organ structures
Input Example			I was running my spellchecker and totally didn't realize that this was a vandalized page. Please accept my apology. I will spellcheck a little slower next time.	This was potentially the most dangerous stunt I have ever seen someone do. One minor mistake and you die.		
Examples of Expert Features			Categories: First person, Please, Negative, Promise, Apologetic "I was running my spellchecker and totally didn't realize that this was a vandalized page. Please accept my apology. I will spellcheck a little slower next time."	"This was potentially the most dangerous stunt I have ever seen someone do. One minor mistake and you die." Low arousal High arousal, negative valence Low arousal, negative valence Positive valence		
Adapted From	[Kacprzak et al., 2023]	[Team et al., 2018]	[Havaldar et al., 2023a]	[Demszky et al., 2020]	[Majkowska et al., 2020]	[Madani et al., 2022]

Figure 1: The FIX benchmark contains 6 datasets across a diverse set of application areas, data modalities, and dataset sizes. For each dataset, we show an example of an input and some example expert features for that input.

they should encompass a grouping of related low-level features (e.g., pixels, tokens), thus creating high-level features that experts can more easily digest. Second, these low-level feature groupings should align with domain experts' knowledge of the relevant task, thus creating features with practical relevance. We refer to features that satisfy these criteria as **expert features.**

But how can we obtain such features? In practice, it is left to domain experts to identify and provide such features for individual tasks. Although experts often have a sense of what the expert features should be, formalizing such features is often non-trivial. Moreover, manually annotating expert features can also be expensive and labor-intensive. These challenges raise the critical question:

Can we automatically discover expert features that align with domain knowledge?

To this end, we present the FIX benchmark, a unified evaluation measuring feature interpretability that can capture each individual domain's expert knowledge. Our goal is to guide the development of new methods that produce interpretable features by building a unified metric to measure how interpretable a proposed feature group is. The FIX datasets (summarized in Figure 1) collectively encompass a diverse array of real-world settings (cosmology, psychology, and medicine) and data modalities (vision, language, and time-series signals): abdomen surgery safety identification (Madani et al., 2022), chest X-ray classification (Lian et al., 2021), mass maps regression (Kacprzak et al., 2023), supernova classification (Željko Ivezić et al., 2019), multilingual politeness classification (Havaldar et al., 2023a), and emotion classification (Demszky et al., 2020; Havaldar et al., 2023b). The challenge here lies in unifying all 6 different real-world settings and 3 different data modalities into a *single* framework, which our proposed expert alignment measure FIXSCORE achieves. This allows us to have a benchmark that does not overfit to any particular domain. To our knowledge, while previous work has identified the need for interpretable features (Zytek et al., 2022; Doshi-Velez & Kim, 2017), there does not exist yet a benchmark that measures the interpretability of features for real-world experts. The FIX benchmark accomplishes this while also serving as a basis for studying, constructing, and extracting expert features. In summary, our contributions are as follows:

1. In collaboration with domain experts, we develop the **FIX** benchmark, a set of 6 curated datasets with evaluation metrics for extracting **Features Interpretable to eXperts** in real-world settings from diverse modalities of images, text, and time-series data. [1]

[1] Code and updates are available at `https://anonymous_website.html`

2. We introduce a general feature evaluation metric, FIXSCORE, that unifies the different real-world settings of cosmology, psychology, and medicine into a single framework. We worked closely with real domain experts to develop criteria for what made features interpretable in each domain.
3. We evaluate commonly used techniques for extracting higher-level features and find that existing methods score poorly on FIXSCORE, highlighting the need for developing new general-purpose methods designed to automatically extract expert features.

2 RELATED WORK

Interpretability. Interpretability in machine learning is often viewed as a multifaceted concept that encompasses algorithmic transparency (Shin & Park, 2019; Rader et al., 2018; Grimmelikhuijsen, 2023), explanation methods (Marcinkevičs & Vogt, 2023; Havaldar et al., 2023c), and visualization techniques (Choo & Liu, 2018; Spinner et al., 2019; Wang et al., 2023), among other aspects. In this work, we focus on feature-level interpretability, a central topic in interpretability research (Hong et al., 2020; Nauta et al., 2023). Feature-based methods are popular because they are believed to offer simple, adaptable, and intuitive settings in which to analyze and develop interpretable machine learning workflows (Molnar et al., 2020). We refer to (Nauta et al., 2023; Dwivedi et al., 2023; Weber et al., 2023) and the references therein for extensive reviews on feature-based explanations.

Application-grounded Evaluation. Chaleshtori et al. (2024) extend the work of Doshi-Velez & Kim (2017) to propose a comprehensive taxonomy of evaluating explanations. Notably, this includes *application-grounded evaluations*, which broadly seek to measure the efficacy of feature-based methods in settings with human users and realistic tasks, such as AI-assisted decision-making. However, the available literature on application-grounded evaluations is sparse: Chaleshtori et al. (2024) reviewed over 50 existing NLP datasets and found that only four were suitable for application-grounded evaluations (DeYoung et al., 2019; Wadden et al., 2020; Koreeda & Manning, 2021; Malik et al., 2021). A principal objective of the FIX benchmark is to provide an application-grounded evaluation of feature-based explanations in real-world settings.

Feature Generation. Because high-quality and interpretable features may not always be available, there is interest in automatically generating them by combining low-level features (Nargesian et al., 2017; Erickson et al., 2020; Zhang et al., 2023a). Notably, Zhang et al. (2023a) propose a method for tabular data using the expand-and-reduce framework (Kanter & Veeramachaneni, 2015). However, existing generation methods do not necessarily produce interpretable features, and most works focus on tabular data. The FIX benchmark aims to address these limitations by providing a setting in which to study and develop methods for interpretable feature generation across diverse problem domains.

XAI Benchmarks. There exists a suite of benchmarks for explanations that cover the properties of faithfulness (or fidelity) (Zhou et al., 2021; Agarwal et al., 2022), robustness (Alvarez-Melis & Jaakkola, 2018; Agarwal et al., 2022), simulatability (Mills et al., 2023), fairness (Fel et al., 2021; Agarwal et al., 2022), among others. Quantus (Hedström et al., 2023), XAI-Bench (Liu et al., 2021), OpenXAI (Agarwal et al., 2022), GraphXAI (Agarwal et al., 2023), and ROAR (Hooker et al., 2019) are notable open-source implementations that evaluate for such properties. CLEVR-XAI (Arras et al., 2022) and Zhang et al. (2023b) provide benchmarks that combine vision and text. ERASER (DeYoung et al., 2019) is a popular NLP benchmark that unifies diverse NLP datasets of human rationales and decisions. In general, however, there is a lack of interpretability benchmarks that evaluate feature interpretability in real-world settings — a gap we aim to address with the FIX benchmark.

3 EXPERT FEATURE EXTRACTION

Feature-based explanation methods require interpretable features to be effective. For example, surgeons communicate safety in surgery with respect to key anatomical structures and organs, which are interpretable features for surgeons (Strasberg & Brunt, 2010; Hashimoto et al., 2019). These interpretable features are a key bridge that can help surgical AI assistants communicate effectively with surgeons. However, ground-truth annotations for such interpretable features are often expensive and hard to obtain, as they typically require trained experts to manually annotate large amounts of data. This bottleneck is not unique to surgery, and such challenges motivate us to study the problem of extracting *features interpretable to experts*, or expert features.

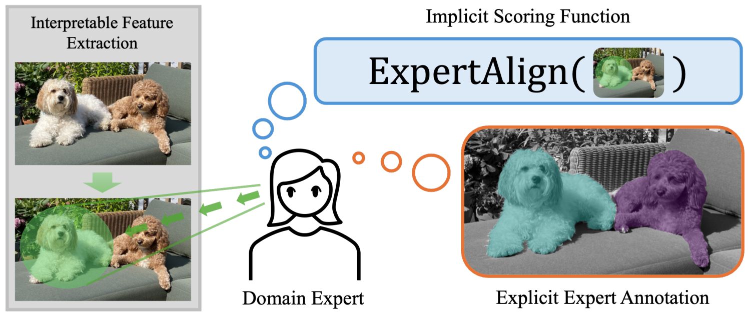


Figure 2: The FIX benchmark allows measuring alignment of extracted features with expert features in different domains, either implicitly with a scoring function or explicitly with expert annotations.

Consider a task with inputs from $\mathcal{X} \subseteq \mathbb{R}^d$ and outputs in $\mathcal{Y}$. In the example of surgery, $\mathcal{X}$ may be the set of surgery images, and $\mathcal{Y}$ is the target of where it is safe or unsafe to operate. We model a higher-level expert feature of input $x \in \mathcal{X}$ as a subset of features represented with a binary mask $g \in \{0,1\}^d$, where $g_i = 1$ if the ith feature is included and 0 otherwise. In surgery, for example, a good mask β is one that accurately selects a key anatomical structure or organ from an input x. The objective of interpretable feature extraction is to find a set of masks $\hat{G} \subseteq \{0,1\}^d$ that effectively approximates the expert features of x. That is, each binary mask $\hat{g} \in \hat{G}$ aims to identify some subset of features meaningful to experts.

3.1 Measuring Alignment of Extracted Features with Expert Features

At the core of FIX is a general framework for measuring the quality of extracted features with respect to expert knowledge. Let $\hat{G}$ be a proposed set of expert features for an input $x \in \mathbb{R}^d$, and suppose there exists a function $\textsc{ExpertAlign}(\hat{g}, x) \in [0,1]$ that captures how well a single extracted feature $\hat{g}$ is expert-interpretable for x. Here, a score of 1 means that a domain expert considers $\hat{g}$ highly interpretable, whereas a score of 0 means that $\hat{g}$ is a highly uninterpretable feature. Then, given a set of proposed groups $\hat{G}$ and input x, we measure the quality of $\hat{G}$ for x as:

$$\text{FIX}\textsc{Score}(\hat{G}, x) = \frac{1}{d} \sum_{i=1}^{d} \frac{1}{|\hat{G}[i]|} \sum_{\hat{g} \in \hat{G}[i]} \textsc{ExpertAlign}(\hat{g}, x). \tag{1}$$

where let $\hat{G}[i] = \{\hat{g} \in \hat{G} : i \in \hat{g}\}$ be the subset $\hat{G}$ that cover feature i. Intuitively, FIXSCORE is an average of averages: the expert alignment for each individual feature $i = 1, \dots, d$ is averaged over all covers $\hat{G}[i]$. This metric has two key strengths:

1. **Duplication Invariance at Optimality.** If one extracts perfect expert features (i.e., with an alignment score of 1), the FIXSCORE cannot be increased further by duplicating expert features. This property ensures that the score cannot be trivially inflated with repeats.
2. **Encourages Diversity of Expert Features.** Since the score aggregates a value for each feature from $i = 1, \dots, d$, adding a new expert feature that does not yet overlap with already extracted features is always beneficial.

The use of a generic expert alignment function enables the FIXSCORE to accommodate a diverse set of applications. There are two main ways one can specify the EXPERTALIGN function: implicitly with a score specified by an expert or explicitly with annotations from an expert, as shown in Figure 2.

Case 1: Implicit Expert Alignment. Suppose we do not have explicit annotations of expert features for ground truth groups. In this case, we use implicit expert features defined indirectly via a scoring function that measures the quality of an extracted feature. The exact formula of the score is specified by an expert and will depend on the domain and task. Implicit expert features have the advantage of potentially being more scalable than features manually annotated by experts. The Mass Maps, Supernova, Multilingual Politeness, and Emotion datasets are examples of the implicit features case.

Case 2: Explicit Expert Alignment. In the case where we do have annotations for expert features $G^\star$, we can use a standardized expression for the FIXSCORE that measures the best possible intersection with the annotated expert features. Then, the expert alignment score of a feature group $\hat{g}$ is

$$\textsc{ExpertAlign}(\hat{g}, x) = \max_{g^\star \in G^\star(x)} \textsc{Match}(\hat{g}, g^\star), \quad \text{where } \textsc{Match}(\hat{g}, g^\star) = \frac{|\hat{g} \cap g^\star|}{|\hat{g} \cup g^\star|}, \tag{2}$$

and $|\cdot|$ counts the number of ones-entries, and $\cap$ and $\cup$ are the element-wise conjunction and disjunction of two binary vectors, respectively. In other words, MATCH is an intersection-over-union score. Our notation is motivated by the fact that one can treat expert features $\hat{g}$ like sets as they are binary vectors. The Cholecystectomy and Chest X-ray datasets have explicit expert features.

Our goal in FIX is to benchmark general-purpose feature extraction techniques that are *domain agnostic* and do not use the FIXSCORE during training. Instead, benchmark challengers can use neural network models trained on the end-to-end tasks to automatically extract features without explicit supervision, which we release as part of the benchmark and discuss further in Appendix B. Annotations for expert features are too expensive to collect at scale for training, while implicit features are by no means comprehensive. The FIX benchmark is intended for evaluation purposes to spur research in general purpose and automated expert feature extraction.

4 FIX DATASETS

In this section, we briefly describe each FIX dataset in Figure 1. For each dataset, we provide an overview of the domain task and the problem setup. We then introduce the key expert alignment function that measures the quality of an expert feature, and explain why certain properties incorporated in the expert alignment function are desirable to experts.

4.1 MASS MAPS DATASET

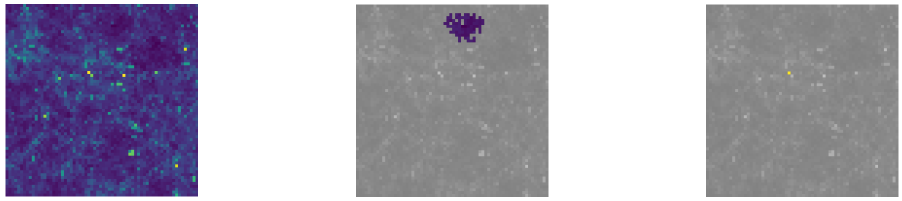
Motivation. A major focus of cosmology is the initial state of the universe, which can be characterized by various cosmological parameters such as Ω_m, which relates to energy density, and σ_8, which pertains to matter fluctuations (Abbott et al., 2022). These parameters influence what is observable by mass maps, also known as weak lensing maps, which capture the spatial distribution of matter density in the universe. Although mass maps can be obtained through the precise measurement of galaxies (Jeffrey et al., 2021; Gatti et al., 2021), it is not known how to directly measure Ω_m and σ_8. This has inspired machine learning efforts to predict the two cosmological parameters from simulations (Ribli et al., 2019; Matilla et al., 2020; Fluri et al., 2022). However, it is hard for cosmologists to gain insights into how to predict Ω_m and σ_8 from black-box ML models.

Problem Setup. Our dataset contains clean simulations from CosmoGridV1 (Kacprzak et al., 2023). Each input is a one-channel image of size $(66, 66)$, where the task is to predict Ω_m and σ_8. Here, Ω_m is the average energy density of all matter relative to the total energy density, including radiation and dark energy, while σ_8 describes fluctuations in the distribution of matter (Abbott et al., 2022). The dataset has 90,000/10,000/10,000 mass maps in train/validation/test splits.

Expert Features. When inferring Ω_m and σ_8 from the mass maps, we aim to discover which cosmological structures most influence these parameters. Two types of cosmological structures in mass maps known to cosmologists are voids and clusters (Matilla et al., 2020). An example is illustrated in Figure 3, where voids are large regions that are under-dense relative to the mean density and appear as dark, while clusters are over-dense and appear as bright dots.

To quantify the interpretability of an expert feature in the mass maps, we develop an implicit expert alignment scoring function. Intuitively, a group that is purely void or purely cluster is more interpretable in cosmology, while a group that is a mixture is less interpretable. We thus develop the purity metric based on the entropy among void/cluster pixels (Zhang et al., 2003) weighted by the ratio of interpretable pixels in the expert feature. We give additional details in Appendix A.1.

$$\textsc{ExpertAlign}(\hat{g}, x) = \text{Purity}_{vc}(\hat{g}, x) \cdot \text{Ratio}_{vc}(\hat{g}, x) \tag{3}$$



(a) Full map (b) Void (c) Cluster

Figure 3: An example with expert features for Mass Maps Regression, showing (a) the full map, (b) a feature with 100% void, and (c) a feature with 100% cluster. Voids are under-dense large regions that appear to be dark, and clusters are over-dense regions that appear as bright dots. The purity scores for both void and cluster are 1. We gray-out the pixels not selected in each feature.

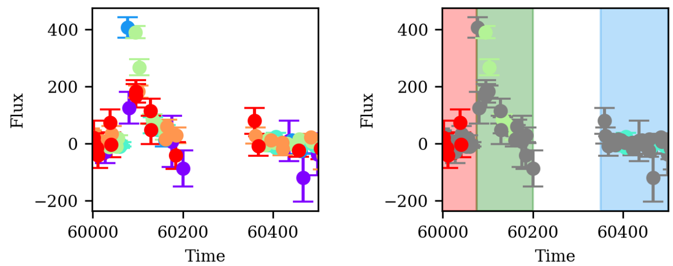


Figure 4: An example with expert features for supernova classification, showing (left) the original time-series dataset and (right) an example of the interpretable expert feature group. We highlight the expert feature groups with the highest expert align scores.

4.2 Supernova Dataset

Motivation. The astronomical time-series classification, as mentioned in (Team et al., 2018), involves categorizing astronomical sources that change over time. Astronomical sources include transient phenomena (e.g. supernovae, kilonovae) and variable objects (e.g. active galactic nuclei, Mira variables). This task analyzes simulation datasets that emulate future telescope observations from the Legacy Survey of Space and Time (LSST) (Željko Ivezić et al., 2019). Given the vastness of the universe, it is essential to identify the time periods that have the most significant impact on classification of astronomical sources to optimize telescope observations. Time periods with no observed data are less useful. To avoid costly searching over all timestamps for high-influence time periods, we aim to identify significant timestamps that are linearly consistent in specific wavelengths.

Problem Setup. We take parts of the dataset from the original PLAsTiCC challenge (Team et al., 2018). The input data are simulated LSST observations comprising four columns: observation times (modified Julian days), wavelength (filter), flux values, and flux error. The dataset encompasses 7 distinct wavelengths that work as filters, and the flux values and errors are recorded at specific time intervals for each wavelength. The classification task is to predict whether or not each of 14 different astronomical objects exists. The supernova dataset contains 6274/728/792 train/valid/test examples.

Expert Features. A feature with linearly consistent flux for each wavelength is considered more interpretable in astrophysics. An illustration of expert features used for supernova classification is presented in Figure 4. This example showcases the flux value and error for various wavelengths, each represented by a different color. We colored the timestamp of expert features with the wavelength color with the highest linear consistency score. For the timestamp where there is no data point, we do not recognize it as an expert feature. We create a linear consistency metric to assess the expert alignment score of a proposed feature in the context of a supernova. Our linear consistency metric uses p, the percentage of data points that display linear consistency, penalized by d, the percentage of time stamps containing data points:

$$\textsc{ExpertAlign}(\hat{g}, x) = \max_{w \in W} p(\hat{g}, x_w) \cdot d(\hat{g}, x_w). \tag{4}$$

where W is the set of unique wavelength. Further details are provided in Appendix A.2.

Example	Expert Features with High Alignment
[Politeness] I was running my spellchecker and totally didn't realize that this was a vandalized page. Please accept my apology. I will spellcheck a little slower next time.	g_1 = `I, my, I` g_2 = `spellchecker, vandalized, little, slower` g_3 = `will` g_4 = `my, apology`
[Emotion] This was potentially the most dangerous stunt I have ever seen someone do. One minor mistake and you die.	g_1 = `dangerous, die` g_2 = `potentially, minor` g_3 = `mistake, stunt` g_4 = `I, someone, you`

Table 1: Examples and expert features with high expert alignment for Multilingual Politeness (top) and Emotion (bottom). These expert features correspond to low distance within the emotion circumplex and high similarity with politeness lexica, respectively.

4.3 Multilingual Politeness Dataset

Motivation. Different cultures express politeness differently (Leech, 2007; Pishghadam & Navari, 2012). For instance, politeness in Japan often involves acknowledging the place of others (Spencer-Oatey & Kádár, 2016), whereas politeness in Spanish-speaking countries focuses on establishing mutual respect (Placencia & Garcia-Fernandez, 2017). Therefore, grounding interpretable features that indicate politeness is *language-dependent*. Previous work from Danescu-Niculescu-Mizil et al. (2013) and Li et al. (2020) use past politeness research to create lexica that indicate politeness/rudeness in English and Chinese, respectively. A lexicon is a set of categories where each category contains a curated list of words. For instance, the English politeness lexicon contains categories like *Gratitude*: "appreciate", "thank you", et cetera, and *Apologizing*: "sorry", "apologies", etc. Havaldar et al. (2023a) expand on these theory-grounded lexica to include Spanish and Japanese.

Problem Setup. The multilingual politeness dataset from (Havaldar et al., 2023a) contains 22,800 conversation snippets from Wikipedia's editor talk pages. The dataset spans English, Spanish, Chinese, and Japanese, and native speakers of these languages have annotated each conversation snippet for politeness level, ranging from -2 (very rude) to 0 (neutral) to 2 (very polite).

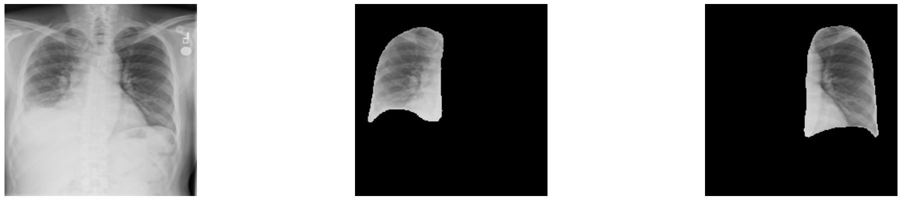
Expert Features. When extracting interpretable features for a task like politeness classification across multiple languages, it is useful to ground these features using prior research from communication and psychology. If extracted politeness features from an LLM are interpretable and domain-aligned, they should match what psychologists have determined to be key politeness indicators. Examples of expert-aligned features are shown in Table 1. Concretely, for each lexical category, we use an LLM to embed all the contained words and then average the resulting embeddings to get a set C of k centroids: $C = c_1, c_2, ...c_k$. See Appendix A.3 for more details. Then, a proposed expert feature $\hat{g} \in \{0, 1\}^d$ indicates whether or not each of the d words $w_1, w_2, ..., w_d \in x$ are included in the feature, and the expert alignment score for the proposed feature $\hat{g}$ can be computed as follows:

$$\textsc{ExpertAlign}(\hat{g}, x) = \max_{c \in C} \frac{1}{|\hat{g}|} \sum_{i=1}^{d} \hat{g}_i \cdot \cos(\text{embedding}(w_i), c) \tag{5}$$

4.4 Emotion Dataset

Motivation. Emotion classification involves inferring the emotion (e.g., Joy, Anger, etc.) reflected in a piece of text. Researchers study emotion to build systems that can understand emotion and thus adapt accordingly when interacting with human users. For extracted features to be useful for such systems, they must be relevant to emotion. For example, a word like "puppy" may be used more frequently in comments labeled with Joy vs. other emotions; therefore, it may be extracted as a relevant feature for the Joy class. However, this is a spurious correlation — emotional expression is not necessarily tied to a subject, and comments containing "puppy" may also be angry or sad.

Problem Setup. The GoEmotions dataset from Demszky et al. (2020) contains 58,000 English Reddit comments labeled for 27 emotion categories, or "neutral" if no emotion is applicable. The



(a) Full image (b) Right lung (c) Left lung

Figure 5: An example with expert features for Chest X-Ray dataset. (a) The full X-ray image where the following pathologies are present: effusion, infiltration, and pneumothorax; (b-c) Expert-interpretable anatomical structures of the left and right lungs.

input is a text utterance of 1-2 sentences extracted from Reddit comments, and the output is a binary label for each of the 27 emotion categories.

Expert Features. Example expert features are shown in Table 1. To measure how emotion-related a feature is, we use the circumplex model of affect (Russell, 1980). The circumplex model assumes that all emotions can be projected onto the 2D unit circle with respect to two independent dimensions – *arousal* (the magnitude of intensity or activation) and *valence* (how negative or positive). By projecting features onto the unit circle, we can quantify emotional relations. In particular, we calculate the following two attributes of the features with a group: (1) their emotional *signal*, i.e., mean distance to the circumplex and (2) their emotional *relatedness*, i.e., mean pairwise distance within the circumplex. We then calculate the following: $\text{Signal}(\hat{g}, x)$, which measures the average Euclidean distance to the circumplex for every projected feature in $\hat{g}$, and $\text{Relatedness}(\hat{g}, x)$, which measures the average pairwise distance between every projected feature in $\hat{g}$ (details in Appendix A.4). For an extracted feature $\hat{g}$, the expert alignment score can then be computed by:

$$\textsc{ExpertAlign}(\hat{g}, x) = \tanh(\exp[-\text{Signal}(\hat{g}, x) \cdot \text{Relatedness}(\hat{g}, x)]) \tag{6}$$

4.5 Chest X-Ray Dataset

Motivation. Chest X-ray imaging is a common procedure for diagnosing conditions such as atelectasis, cardiomegaly, and effusion, among others. Although radiologists are skilled at analyzing such images, modern machine learning models are increasingly competitive in diagnostic performance (Ahmad, 2021). Therefore, ML models may prove useful in assisting radiologists in making diagnoses. However, in the absence of an explanation, radiologists may only trust the model output if it matches their own predictions. Moreover, inaccurate AI assistants are shown to negatively affect diagnostic performance (Yu et al., 2024). To address this problem, explainability could be employed as a safeguard to help radiologists decide whether or not to trust the model. As such, it is important for machine learning models to provide explanations for their diagnoses.

Problem Setup. We use the NIH-Google dataset (Majkowska et al., 2020) available from the TorchXRayVision library (Cohen et al., 2022). This is a relabeling of the NIH ChestX-ray14 dataset (Wang et al., 2017) which improved the quality of the original labels. It contains 28,868 chest X-ray images labeled for 14 common pathology categories: atelectasis, calcification, cardiomegaly, etc. We randomly partition the dataset into train/test splits of 23,094 and 5,774, respectively. The task is a multi-label classification problem for identifying the presence of each pathology.

Expert Features. Radiology reports commonly refer to anatomical structures (e.g., spine, lungs), which allows radiologists to perform and communicate accurate diagnoses to patients. We provide these expert-interpretable features in the form of anatomical structure segmentations. However, because we could not find datasets with both pathology labels and anatomical segmentations, we used a pre-trained model from TorchXRayVision to generate the structure labelings for each image. We use explicit expert alignment as described in Equation 2 to compute alignment of an extracted feature $\hat{g}$ and the 14 predicted anatomical structure segments, including the left clavicle, heart, etc. Details of the Chest X-Ray dataset can be found in Appendix A.5.

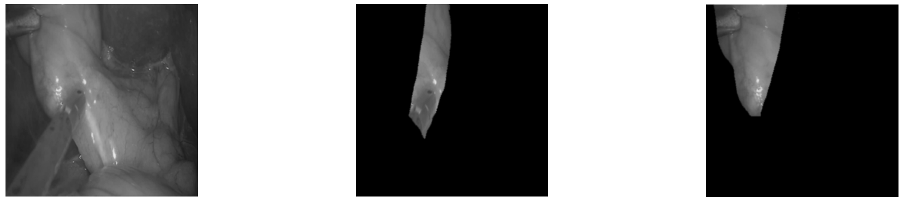


Figure 6: An example with expert features of Laparoscopic Cholecystectomy Surgery Dataset: (a) The view of the surgeon sees; (b) The safe region for operations; (c) The gallbladder, a key anatomical structure for the critical view of safety.

4.6 Laparoscopic Cholecystectomy Surgery Dataset

Motivation. Laparoscopic cholecystectomy (gallbladder removal) is one of the most common elective abdominal surgeries performed in the US, with over 750,000 operations annually (Stinton & Shaffer, 2012). A common complication of laparoscopic surgery is bile duct injury, which is associated with an 8-fold increase in mortality (Michael Brunt et al., 2020) and accounts for more than $1B in US healthcare annual spending (Berci et al., 2013). Notably, 97% of such complications result from human visualization errors (Way et al., 2003). The surgery site commonly contains obstructing tissues, inflammation, and other patient-specific artifacts — all of which may prevent the surgeon from getting a perfect view. Consequently, there is growing interest in harnessing advanced vision models to help surgeons distinguish safe and risky areas for operation. However, experienced surgeons rarely trust model outputs due to their opaque nature, while inexperienced surgeons might overly rely on model predictions. Therefore, any safe and useful machine learning model must be able to provide explanations that align with surgeons' expectations.

Problem Setup. The task is to identify the safe and unsafe regions for incision. We use the open-source subset of the data from (Madani et al., 2022), wherein the authors enlist surgeons to annotate surgery video data from the M2CAI16 workflow challenge (Stauder et al., 2016) and Cholec80 (Twinanda et al., 2016) datasets. This dataset consists of 1015 annotated images with a random train/test split of 812 and 203, respectively.

Expert Features. In cholecystectomy, it is a common practice for surgeons to identify the *critical view of safety* before performing any irreversible operations (Strasberg & Brunt, 2010; Hashimoto et al., 2019). This view identifies the location of vital organs and structures that inform the safe region of operation and is incidentally what surgeons often need as part of an explanation. We provide these expert-interpretable labels in the form of organ segmentations (liver, gallbladder, hepatocystic triangle). We use explicit expert alignment as described in Equation 2 to compute alignment of an extracted feature $\hat{g}$ and the surgeon-annotated organ labels taken from Madani et al. (2022). Details of the Cholecystectomy dataset can be found in Appendix A.6.

5 Baseline Algorithms & Discussion

We evaluate standard techniques widely used within the vision, text, and time series domains to create higher-level features. We provide a brief summary below, with additional details in Appendix C.

Domain-specific Baselines. We consider the following domain-centric baselines. *(Image)* For image data, we consider three segmentation methods (Kim et al., 2024). Patches (Dosovitskiy et al., 2021) divides the image into grids where each cell is the same size. Quickshift (Grady, 2006) connects similar neighboring pixels into a common superpixel. Watershed (Levner & Zhang, 2007) simulates flooding on a topographic surface. CRAFT (Fel et al., 2023) generates concept attribution maps. *(Time-series)* For time series data, we take equal size slices of the data across time as patches (Schlegel et al., 2021). We use different slice sizes to see how they impact multiple baselines. We take various slice sizes, such as 5, 10, and 15, separately to evaluate the results of multiple baselines. *(Text)* For text data, we present three baselines for extracting features (Rychener et al., 2022). At the finest granularity, we treat each word as a feature. The second baseline considers each phrase as a feature.

	Vision				Time Series		Language		
	Method	**Cholec**	**ChestX**	**MassMaps**	**Method**	**Supernova**	**Method**	**Politeness**	**Emotion**
Domain-specific	Identity	0.4686	0.2154	0.5486	Identity	0.0152	Identity	0.6070	0.0103
	Random	0.1086	0.0427	0.5508	Random	0.0358	Random	0.6478	0.0303
	Patch	0.0323	0.0999	0.5549	Slice 5	0.0337	Words	0.6851	0.1182
	Quickshift	0.2622	0.3419	0.5496	Slice 10	0.0555	Phrases	0.6351	0.0198
	Watershed	0.2807	0.1452	0.5594	Slice 15	0.0550	Sentences	0.6109	0.0120
	SAM	0.3678	0.3151	0.5526					
	CRAFT	0.0271	0.1175	0.3991					
Domain-agnostic	Clustering	0.2880	0.2627	0.5518	Clustering	0.2622	Clustering	0.6680	0.0912
	Archipelago	0.3351	0.2148	0.5509	Archipelago	0.2574	Archipelago	0.6773	0.0527

Table 2: Baselines scores of different FIX settings. We report the mean score and give a more comprehensive table in Appendix C. We describe baseline implementations in Section 5. One thing to note is that FIXSCORE is not comparable for different tasks (e.g. between Mass Maps and Supernova) as the data and specific expert alignment metrics are different for different tasks.

Phrases are comprised of groups of words that are separated by some punctuation in the original text. At the coarsest granularity, we treat each sentence as a feature.

Domain-agnostic Baselines. We additionally consider the following domain-agnostic baselines for feature extraction. *(Identity)* We combine all elements into one single group. *(Random)* We select features at random, up to the maximum baseline results for the group. The group maximum is calculated as: (group maximum) $\approx$ (scaling factor) $\times$ (number of expert features). The size of the distinct expert feature varies depending on the setting, and further details for each setting can be found in Appendix C. We use a scaling factor of about 1.5 to allow for flexibility. *(Clustering)* For images, we first use Quickshift to generate segments and then pass each segment through a feature extractor (ResNet-18 by default). For time series, we use raw features from each time segment. We then apply K-means clustering on the extracted/raw features to relabel and merge segments. For text, we use BERTopic (Grootendorst, 2022) to obtain the clusters. *(Archipelago)* We adapt the implementation of Archipelago (Tsang et al., 2020) to use ResNet-18 with quickshift for feature extraction.

Results and Discussions. We show results on the baselines in Table 2. For image datasets, Quickshift has the best performance compared to Patch and Watershed on both the Cholecystectomy dataset and the Chest X-ray dataset, since they have natural images. All baselines perform similarly for the Mass Maps dataset. That the range of mass maps is different from other tasks is potentially because they are not natural images, but rather similar to topographic surfaces. For the Supernova time-series dataset, larger slices score yield higher expert alignment scores. For both Multilingual Politeness and Emotion datasets, individual words appear to be the most expert-aligned features. Generally, however, we see that the domain-agnostic neural baselines tend to also perform better than or close to the best domain-centric baseline. The main benefit of using a neural approach is that it can more easily automatically discover relevant features.

6 CONCLUSION

We propose FIX, a curated benchmark of datasets with evaluation metrics for extracting expert features in diverse real-world settings. Our benchmark addresses a gap in the literature by providing researchers with an environment to study and automatically extract interpretable features for experts.

Limitations and Future Work. The FIX benchmark is not an exhaustive specification of all expert features, and may fail to capture others types. The ones we included are generally non-controversial and well-accepted by the domain's expert community, but we can foresee that there are cases where this may not be true. Dealing with potential conflicting expert opinions may need a more nuanced approach, which is left for future work to address. Furthermore, although we cover cosmology, psychology, and medicine domains in this work, the metrics for these domains may not be appropriate for all settings. We encourage prospective users to consider and implement metrics most appropriate to their particular settings. Future work includes the development of new, general purpose techniques that can extract expert features from data and models without supervision.

Reproducibility Statement. Our code is open-source at `https://anonymous_website.com`.

Ethics Statement. This work seeks to make explainable machine learning more accessible to experts. However, like the ML models, explanation methods are fallible and therefore should still be regarded thoughtfully by users.

REFERENCES

T. M.C. Abbott, M. Aguena, A. Alarcon, S. Allam, O. Alves, A. Amon, F. Andrade-Oliveira, J. Annis, S. Avila, D. Bacon, E. Baxter, K. Bechtol, M.R. Becker, G.M. Bernstein, S. Bhargava, S. Birrer, J. Blazek, A. Brandao-Souza, S.L. Bridle, D. Brooks, E. Buckley-Geer, D.L. Burke, H. Camacho, A. Campos, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F.J. Castander, R. Cawthon, C. Chang, A. Chen, R. Chen, A. Choi, C. Conselice, J. Cordero, M. Costanzi, M. Crocce, L.N. da Costa, M.E. da Silva Pereira, C. Davis, T.M. Davis, J. De Vicente, J. DeRose, S. Desai, E. Di Valentino, H.T. Diehl, J.P. Dietrich, S. Dodelson, P. Doel, C. Doux, A. Drlica-Wagner, K. Eckert, T.F. Eifler, F. Elsner, J. Elvin-Poole, S. Everett, A.E. Evrard, X. Fang, A. Farahi, E. Fernandez, I. Ferrero, A. Ferté, P. Fosalba, O. Friedrich, J. Frieman, J. García-Bellido, M. Gatti, E. Gaztanaga, D.W. Gerdes, T. Giannantonio, G. Giannini, D. Gruen, R.A. Gruendl, J. Gschwend, G. Gutierrez, I. Harrison, W.G. Hartley, K. Herner, S.R. Hinton, D.L. Hollowood, K. Honscheid, B. Hoyle, E.M. Huff, D. Huterer, B. Jain, D.J. James, M. Jarvis, N. Jeffrey, T. Jeltema, A. Kovacs, E. Krause, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, P.-F. Leget, P. Lemos, A.R. Liddle, C. Lidman, M. Lima, H. Lin, N. MacCrann, M.A.G. Maia, J.L. Marshall, P. Martini, J. McCullough, P. Melchior, J. Mena-Fernández, F. Menanteau, R. Miquel, J.J. Mohr, R. Morgan, J. Muir, J. Myles, S. Nadathur, A. Navarro-Alsina, R.C. Nichol, R.L.C. Ogando, Y. Omori, A. Palmese, S. Pandey, Y. Park, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas Malagón, A. Porredon, J. Prat, M. Raveri, M. Rodriguez-Monroy, R.P. Rollins, A.K. Romer, A. Roodman, R. Rosenfeld, A.J. Ross, E.S. Rykoff, S. Samuroff, C. Sánchez, E. Sanchez, J. Sanchez, D. Sanchez Cid, V. Scarpine, M. Schubnell, D. Scolnic, L.F. Secco, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. Smith, M. Soares-Santos, E. Suchyta, M.E.C. Swanson, M. Tabbutt, G. Tarle, D. Thomas, C. To, A. Troja, M.A. Troxel, D.L. Tucker, I. Tutusaus, T.N. Varga, A.R. Walker, N. Weaverdyck, R. Wechsler, J. Weller, B. Yanny, B. Yin, Y. Zhang, and J. Zuntz and. Dark energy survey year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2), January 2022. doi: 10.1103/physrevd.105.023520. URL `https://doi.org/10.1103%2Fphysrevd.105.023520`.

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in neural information processing systems*, 35:15784–15799, 2022.

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, 2023.

Rani Ahmad. Reviewing the relationship between machines and radiology: the application of artificial intelligence. *Acta Radiologica Open*, 10(2):2058460121990296, 2021.

Tarek Allam Jr, Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluís Galbany, Renée Hložek, Emille EO Ishida, Saurabh W Jha, David O Jones, Richard Kessler, et al. The photometric lsst astronomical time-series classification challenge (plasticc): Data set. *arXiv preprint arXiv:1810.00001*, 2018. URL `https://kaggle.com/competitions/PLAsTiCC-2018`.

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. URL `http://arxiv.org/abs/1806.08049`.

Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.11.008. URL `https://www.sciencedirect.com/science/article/pii/S1566253521002335`.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019.

Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. Explanation in ai and law: Past, present and future. *Artificial Intelligence*, 289:103387, 2020. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2020.103387. URL `https://www.sciencedirect.com/science/article/pii/S0004370220301375`.

George Berci, John Hunter, Leon Morgenstern, Maurice Arregui, Michael Brunt, Brandon Carroll, Michael Edye, David Fermelia, George Ferzli, Frederick Greene, et al. Laparoscopic cholecystectomy: first, do no harm; second, take care of bile duct stones, 2013.

Fateme Hashemi Chaleshtori, Atreya Ghosal, Alexander Gill, Purbid Bambroo, and Ana Marasović. On evaluating explanation utility for human-ai decision making in nlp. *arXiv preprint arXiv:2407.03545*, 2024.

Jaegul Choo and Shixia Liu. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92, 2018.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022. URL `https://github.com/mlmed/torchxrayvision`.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*, 2013.

José de la Torre-López, Aurora Ramírez, and José Raúl Romero. Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10):2171–2194, May 2023. ISSN 1436-5057. doi: 10.1007/s00607-023-01181-x. URL `http://dx.doi.org/10.1007/s00607-023-01181-x`.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL `https://aclanthology.org/2020.acl-main.372`.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data, 2020.

Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *CoRR*, abs/2112.04417, 2021. URL `https://arxiv.org/abs/2112.04417`.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023.

Janis Fluri, Tomasz Kacprzak, Aurelien Lucchi, Aurel Schneider, Alexandre Refregier, and Thomas Hofmann. Full wCDM analysis of KiDS-1000 weak lensing maps using deep learning. *Physical Review D*, 105(8), April 2022. doi: 10.1103/physrevd.105.083518. URL `https://doi.org/10.1103%2Fphysrevd.105.083518`.

M. Gatti, E. Sheldon, A. Amon, M. Becker, M. Troxel, A. Choi, C. Doux, N. MacCrann, A. Navarro-Alsina, I. Harrison, D. Gruen, G. Bernstein, M. Jarvis, L. F. Secco, A. Ferté, T. Shin, J. McCullough, R. P. Rollins, R. Chen, C. Chang, S. Pandey, I. Tutusaus, J. Prat, J. Elvin-Poole, C. Sanchez, A. A. Plazas, A. Roodman, J. Zuntz, T. M. C. Abbott, M. Aguena, S. Allam, J. Annis, S. Avila, D. Bacon, E. Bertin, S. Bhargava, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, T. M. Davis, J. De Vicente, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. Drlica-Wagner, K. Eckert, S. Everett, I. Ferrero, J. Frieman, J. García-Bellido, D. W. Gerdes, T. Giannantonio, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, E. M. Huff, D. Huterer, B. Jain, D. J. James, T. Jeltema, E. Krause, R. Kron, N. Kuropatkin, M. Lima, M. A. G. Maia, J. L. Marshall, R. Miquel, R. Morgan, J. Myles, A. Palmese, F. Paz-Chinchón, E. S. Rykoff, S. Samuroff, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, I. Sevilla-Noarbe, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, C. To, D. L. Tucker, T. N. Varga, R. H. Wechsler, J. Weller, W. Wester, and R. D. Wilkinson. Dark energy survey year 3 results: weak lensing shape catalogue. *MNRAS*, 504(3):4312–4336, July 2021. doi: 10.1093/mnras/stab918.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2021.

L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006. doi: 10.1109/TPAMI.2006.233.

Stephan Grimmelikhuijsen. Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, 83(2): 241–262, 2023.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Mohamed Saif Hameed, Simon Laplante, Caterina Masino, Muhammad Khalid, Haochi Zhang, Sergey Protserov, Jaryd Hunter, Pouria Mashouri, Andras Fecso, Michael Brudno, and Amin Madani. What is the educational value and clinical utility of artifcial intelligence for intraoperative and postoperative video analysis? a survey of surgeons and trainees. *Surgical Endoscopy*, 37, 09 2023. doi: 10.1007/s00464-023-10377-3.

Daniel A Hashimoto, C Gustaf Axelsson, Cara B Jones, Roy Phitayakorn, Emil Petrusa, Sophia K McKinley, Denise Gee, and Carla Pugh. Surgical procedural map scoring for decision-making in laparoscopic cholecystectomy. *The American Journal of Surgery*, 217(2):356–361, 2019.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. Comparing styles across languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6775–6791, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.419. URL `https://aclanthology.org/2023.emnlp-main.419`.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion. In Jeremy Barnes, Orphée De Clercq, and Roman Klinger (eds.), *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pp. 202–214, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.wassa-1.19. URL `https://aclanthology.org/2023.wassa-1.19`.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. Topex: Topic-based explanations for model comparison. *arXiv preprint arXiv:2306.00976*, 2023c.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. Building knowledge-guided lexica to model cultural variation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 211–226, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.12. URL https://aclanthology.org/2024.naacl-long.12.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. URL http://jmlr.org/papers/v24/22-0142.html.

Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay (eds.), *Artificial Intelligence in Education*, pp. 154–168, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93843-1.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, May 2020. ISSN 2573-0142. doi: 10.1145/3392878. URL http://dx.doi.org/10.1145/3392878.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks, 2019.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 624–635, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445923. URL https://doi.org/10.1145/3442188.3445923.

N. Jeffrey, M. Gatti, C. Chang, L. Whiteway, U. Demirbozan, A. Kovacs, G. Pollina, D. Bacon, N. Hamaus, T. Kacprzak, O. Lahav, F. Lanusse, B. Mawdsley, S. Nadathur, J. L. Starck, P. Vielzeuf, D. Zeurcher, A. Alarcon, A. Amon, K. Bechtol, G. M. Bernstein, A. Campos, A. Carnero Rosell, M. Carrasco Kind, R. Cawthon, R. Chen, A. Choi, J. Cordero, C. Davis, J. DeRose, C. Doux, A. Drlica-Wagner, K. Eckert, F. Elsner, J. Elvin-Poole, S. Everett, A. Ferté, G. Giannini, D. Gruen, R. A. Gruendl, I. Harrison, W. G. Hartley, K. Herner, E. M. Huff, D. Huterer, N. Kuropatkin, M. Jarvis, P. F. Leget, N. MacCrann, J. McCullough, J. Muir, J. Myles, A. Navarro-Alsina, S. Pandey, J. Prat, M. Raveri, R. P. Rollins, A. J. Ross, E. S. Rykoff, C. Sánchez, L. F. Secco, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. A. Troxel, I. Tutusaus, T. N. Varga, B. Yanny, B. Yin, Y. Zhang, J. Zuntz, T. M. C. Abbott, M. Aguena, S. Allam, F. Andrade-Oliveira, M. R. Becker, E. Bertin, S. Bhargava, D. Brooks, D. L. Burke, J. Carretero, F. J. Castander, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, M. E. S. Pereira, J. De Vicente, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, I. Ferrero, B. Flaugher, P. Fosalba, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, B. Hoyle, B. Jain, D. J. James, M. Lima, M. A. G. Maia, M. March, J. L. Marshall, P. Melchior, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, A. A. Plazas, M. Rodriguez-Monroy, A. Roodman, E. Sanchez, V. Scarpine, S. Serrano, M. Smith, M. Soares-Santos, E. Suchyta, G. Tarle, D. Thomas, C. To, J. Weller, and DES Collaboration. Dark Energy Survey Year 3 results: Curved-sky weak lensing mass map reconstruction. *MNRAS*, 505(3): 4626–4645, August 2021. doi: 10.1093/mnras/stab1495.

Tomasz Kacprzak, Janis Fluri, Aurel Schneider, Alexandre Refregier, and Joachim Stadel. CosmoGridV1: a simulated LambdaCDM theory prediction for map-level cosmological inference. *JCAP*, 2023(2):050, February 2023. doi: 10.1088/1475-7516/2023/02/050.

James Max Kanter and Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE international conference on data science and advanced analytics (DSAA)*, pp. 1–10. IEEE, 2015.

Chaehyeon Kim, Weiqiu You, Shreya Havaldar, and Eric Wong. Evaluating groups of features via consistency, contiguity, and stability. In *The Second Tiny Papers Track at ICLR*, 2024. URL `https://openreview.net/pdf?id=IP2etbIEuC`.

Yuta Koreeda and Christopher D Manning. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*, 2021.

Geoffrey Leech. Politeness: is there an east-west divide? *Journal of Politeness Research*, 2007.

Ilya Levner and Hong Zhang. Classification-driven watershed segmentation. *IEEE Transactions on Image Processing*, 16(5):1437–1445, 2007. doi: 10.1109/TIP.2007.894239.

Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. Studying politeness across cultures using english twitter and mandarin weibo. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415190. URL `https://doi.org/10.1145/3415190`.

Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021.

Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. *CoRR*, abs/2106.12543, 2021. URL `https://arxiv.org/abs/2106.12543`.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Amin Madani, Babak Namazi, Maria S Altieri, Daniel A Hashimoto, Angela Maria Rivera, Philip H Pucher, Allison Navarrete-Welton, Ganesh Sankaranarayanan, L Michael Brunt, Allan Okrainec, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of surgery*, 276(2):363–369, 2022.

Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*, 2021.

Ričards Marcinkevičs and Julia E Vogt. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(3):e1493, 2023.

José Manuel Zorrilla Matilla, Manasi Sharma, Daniel Hsu, and Zoltán Haiman. Interpreting deep learning models for weak lensing. *Physical Review D*, 102(12), December 2020. ISSN 2470-0029. doi: 10.1103/physrevd.102.123506. URL `http://dx.doi.org/10.1103/physrevd.102.123506`.

Albert Meijer and Martijn Wessels. Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, 42(12):1031–1039, 2019. doi: 10.1080/01900692.2019.1575664. URL `https://doi.org/10.1080/01900692.2019.1575664`.

L Michael Brunt, Daniel J Deziel, Dana A Telem, Steven M Strasberg, Rajesh Aggarwal, Horacio Asbun, Jaap Bonjer, Marian McDonald, Adnan Alseidi, Mike Ujiki, et al. Safe cholecystectomy multi-society practice guideline and state-of-the-art consensus conference on prevention of bile duct injury during cholecystectomy. *Surgical endoscopy*, 34:2827–2855, 2020.

Edmund Mills, Shiye Su, Stuart Russell, and Scott Emmons. Almanacs: A simulatability benchmark for language model explainability, 2023.

Ceena Modarres, Mark Ibrahim, Melissa Louie, and John Paisley. Towards explainable deep learning for credit lending: A case study. *arXiv preprint arXiv:1811.06471*, 2018.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 417–431. Springer, 2020.

Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana, Elias B. Khalil, and Deepak Turaga. Learning feature engineering for classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2529–2535, 2017. doi: 10.24963/ijcai.2017/352. URL https://doi.org/10.24963/ijcai.2017/352.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Reza Pishghadam and Safoora Navari. A study into politeness strategies and politeness markers in advertisements as persuasive tools. *Mediterranean Journal of Social Sciences*, 3(2):161–171, 2012.

María Elena Placencia and Carmen Garcia-Fernandez. *Research on politeness in the Spanish-speaking world*. Routledge, 2017.

Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13, 2018.

Abhishek Rai. An explanation of what, why, and how of explainable ai (xai). *Towards Data Science*, November 2019. URL https://towardsdatascience.com/an-explanation-of-what-why-and-how-of-explainable-ai-xai-117d9c441265. Accessed on September 18, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

Dezső Ribli, Bálint Ármin Pataki, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman, and István Csabai. Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the Royal Astronomical Society*, 490(2):1843–1860, 09 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz2610. URL https://doi.org/10.1093/mnras/stz2610.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161, 1980.

Yves Rychener, Xavier Renard, Djamé Seddah, Pascal Frossard, and Marcin Detyniecki. On the granularity of explanations in model agnostic nlp interpretability. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 498–512. Springer, 2022.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks, 2023.

Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110273. URL https://www.sciencedirect.com/science/article/pii/S0950705123000230.

Udo Schlegel, Duy Lam Vo, Daniel A Keim, and Daniel Seebacher. Ts-mule: Local interpretable model-agnostic explanations for time series forecast models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 5–14. Springer, 2021.

Donghee Shin and Yong Jin Park. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98:277–284, 2019.

Helen Spencer-Oatey and Dániel Z Kádár. The bases of (im) politeness evaluations: Culture, the moral order and the east-west debate. *East Asian Pragmatics*, 1(1):73–106, 2016.

Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26(1):1064–1074, 2019.

Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feußner, and Nassir Navab. The tum lapchole dataset for the m2cai 2016 workflow challenge. *arXiv preprint arXiv:1610.09278*, 2016.

Laura M Stinton and Eldon A Shaffer. Epidemiology of gallbladder disease: cholelithiasis and cancer. *Gut and liver*, 6(2):172, 2012.

Steven M Strasberg and Michael L Brunt. Rationale and use of the critical view of safety in laparoscopic cholecystectomy. *Journal of the American College of Surgeons*, 211(1):132–138, 2010.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

The PLAsTiCC Team, Tarek Allam Jr. au2, Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluís Galbany, Renée Hložek, Emille E. O. Ishida, Saurabh W. Jha, David O. Jones, Richard Kessler, Michelle Lochner, Ashish A. Mahabal, Alex I. Malz, Kaisey S. Mandel, Juan Rafael Martínez-Galarza, Jason D. McEwen, Daniel Muthukrishna, Gautham Narayan, Hiranya Peiris, Christina M. Peters, Kara Ponder, Christian N. Setzer, The LSST Dark Energy Science Collaboration, The LSST Transients, and Variable Stars Science Collaboration. The photometric lsst astronomical time-series classification challenge (plasticc): Data set, 2018.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019. URL `http://arxiv.org/abs/1907.07374`.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. In *Advances in Neural Information Processing Systems*, 2020.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

Ruheng Wang, Yi Jiang, Junru Jin, Chenglin Yin, Haoqing Yu, Fengsheng Wang, Jiuxin Feng, Ran Su, Kenta Nakai, Quan Zou, et al. Deepbio: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. *Nucleic acids research*, 51(7):3017–3029, 2023.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Lawrence W Way, Lygia Stewart, Walter Gantert, Kingsway Liu, Crystine M Lee, Karen Whang, and John G Hunter. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Annals of surgery*, 237(4):460–469, 2003.

Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Information Fusion*, 92:154–176, 2023.

Rebecca Wexler. When a computer program keeps you in jail: How computers are harming criminal justice. *The New York Times*, June 2017. URL `https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html`. Opinion.

Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, and Eric Wong. Sum-of-parts models: Faithful attributions for groups of features, 2023.

Feiyang Yu, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar. Heterogeneity and predictors of the effects of ai assistance on radiologists. *Nature Medicine*, pp. 1–13, 2024.

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(3):689–722, September 2016. ISSN 1467-985X. doi: 10.1111/rssa.12227. URL http://dx.doi.org/10.1111/rssa.12227.

Hui Zhang, Jason E Fritts, and Sally A Goldman. Entropy-based objective evaluation method for image segmentation. In *Storage and Retrieval Methods and Applications for Multimedia 2004*, volume 5307, pp. 38–49. SPIE, 2003.

Tianping Zhang, Zheyu Zhang, Zhiyuan Fan, Haoyan Luo, Fengyuan Liu, Qian Liu, Wei Cao, and Jian Li. Openfe: Automated feature generation with expert-level performance, 2023a.

Yifei Zhang, Siyi Gu, James Song, Bo Pan, Guangji Bai, and Liang Zhao. Xai benchmark for visual explanation, 2023b.

Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021. ISSN 2079-9292. doi: 10.3390/electronics10050593. URL https://www.mdpi.com/2079-9292/10/5/593.

Alexandra Zytek, Ignacio Arnaldo, Dongyu Liu, Laure Berti-Equille, and Kalyan Veeramachaneni. The need for interpretable features: Motivation and taxonomy, 2022.

Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L. Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gangler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Garrett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall, Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, David J. Mills, Connor Miraval, Joachim Moeyens, Fred E. Moolekamp, David G. Monet, Marc Moniez, Serge Monkewitz, Christopher Montgomery, Christopher B. Morrison, Fritz Mueller, Gary P. Muller, Freddy Muñoz Arancibia, Douglas R. Neill, Scott P. Newbry, Jean-Yves Nief, Andrei Nomerotski, Martin Nordby, Paul O'Connor, John Oliver, Scot S. Olivier, Knut Olsen, William O'Mullane, Sandra Ortiz, Shawn Osier, Russell E. Owen, Reynald Pain, Paul E. Palecek, John K. Parejko, James B. Parsons, Nathan M. Pease, J. Matt Peterson, John R. Peterson, Donald L. Petravick, M. E. Libby Petrick, Cathy E. Petry, Francesco Pierfederici, Stephen Pietrowicz, Rob Pike, Philip A. Pinto, Raymond Plante, Stephen Plate, Joel P. Plutchak, Paul A. Price, Michael Prouza, Veljko Radeka, Jayadev Rajagopal, Andrew P. Rasmussen, Nicolas Regnault, Kevin A. Reil, David J. Reiss, Michael A. Reuter, Stephen T. Ridgway, Vincent J. Riot, Steve Ritz, Sean Robinson, William Roby, Aaron Roodman, Wayne Rosing, Cecille Roucelle, Matthew R. Rumore, Stefano Russo, Abhijit Saha, Benoit Sassolas, Terry L. Schalk, Pim Schellart, Rafe H. Schindler, Samuel Schmidt, Donald P. Schneider, Michael D. Schneider, William Schoening, German Schumacher, Megan E. Schwamb, Jacques Sebag, Brian Selvy, Glenn H. Sembroski, Lynn G. Seppala, Andrew Serio, Eduardo Serrano, Richard A. Shaw, Ian Shipsey, Jonathan Sick, Nicole Silvestri, Colin T. Slater, J. Allyn Smith, R. Chris Smith, Shahram Sobhani, Christine Soldahl, Lisa Storrie-Lombardi, Edward Stover, Michael A. Strauss, Rachel A. Street, Christopher W. Stubbs, Ian S. Sullivan, Donald Sweeney, John D. Swinbank, Alexander Szalay, Peter Takacs, Stephen A. Tether, Jon J. Thaler, John Gregg Thayer, Sandrine Thomas, Adam J. Thornton, Vaikunth Thukral, Jeffrey Tice, David E. Trilling, Max Turri, Richard Van Berg, Daniel Vanden Berk, Kurt Vetter, Francoise Virieux, Tomislav Vucina, William Wahl, Lucianne Walkowicz, Brian Walsh, Christopher W. Walter, Daniel L. Wang, Shin-Yawn Wang, Michael Warner, Oliver Wiecha, Beth Willman, Scott E. Winters, David Wittman, Sidney C. Wolff, W. Michael Wood-Vasey, Xiuqin Wu, Bo Xin, Peter Yoachim, and Hu Zhan. Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, mar 2019. doi: 10.3847/1538-4357/ab042c. URL https://dx.doi.org/10.3847/1538-4357/ab042c.

A Dataset Details

All datasets and their respective Croissant metadata records and licenses are available on HuggingFace at the following links.

- **Mass Maps**: `https://anonymous_website.html`
- **Supernova**: `https://anonymous_website.html`
- **Multilingual Politeness**: `https://anonymous_website.html`
- **Emotion**: `https://anonymous_website.html`
- **Chest X-Ray**: `https://anonymous_website.html`
- **Laparoscopic Cholecystectomy Surgery**: `https://anonymous_website.html`

A.1 Mass Maps Dataset

Problem Setup. We randomly split the data to consist of 90,000 train and 10,000 validation maps and maintain the original 10,000 test maps. We follow the post-processing procedure in Jeffrey et al. (2021); You et al. (2023) for low-noise maps. Following previous works (Ribli et al., 2019; Matilla et al., 2020; Fluri et al., 2022; You et al., 2023), we use a CNN-based model for predicting Ω_m and σ_8.

Metric. Let $x \in \mathbb{R}^d$ be the input mass map with $d = H \times W$ pixels, and $g \in \{0,1\}^d$ be a boolean mask g that describes which pixels belong to the group, where $g_i = 1$ if the ith pixel belongs to the group, and 0 otherwise.

We can compute the purity score of each group to void and cluster. We say a pixel is a void (underdensed) pixel if its intensity is below 0, and a cluster (overdensed) pixel if its intensity is above $3\sigma(x)$, following previous works (Matilla et al., 2020; You et al., 2023). We first compute the proportion of void pixels and cluster pixels in feature g

$$P_v(g,x) = \frac{\sum_{i=1}^d \mathbb{1}[g_i x_i < 0]}{g^\intercal \mathbf{1}}, \qquad P_c(g,x) = \frac{\sum_{i=1}^d \mathbb{1}[g_i x_i > 3\sigma(x)]}{g^\intercal \mathbf{1}} \tag{7}$$

where $\mathbf{1} \in 1^d$ is the identity matrix, the numerators count the number of underdensed or overdensed pixels, and $g^\intercal \mathbf{1}$ is the number of pixels in the feature. In practice, we add a small $\epsilon = 10^{-6}$ to P_v and P_c and renormalize them, to avoid taking the log of 0 later. Next, we compute the proportion of pixels that are void or cluster, only among the void/cluster pixels:

$$P'_v(g,x) = \frac{P_v(g,x)}{P_v(g,x) + P_c(g,x)}, \qquad P'_c(g,x) = \frac{P_c(g,x)}{P_v(g,x) + P_c(g,x)} \tag{8}$$

Then, we compute the ExpertAlign score for the predicted feature $\hat{g}$ by computing the void/cluster-only entropy reversed and scaled to $[0,1]$, weighted by the percentage of void/cluster pixels among all pixels.

$$\text{Purity}_{vc}(\hat{g},x) = \frac{1}{2}(2 + P'_v(\hat{g},x)\log_2 P'_v(\hat{g},x) + P'_c(\hat{g},x)\log_2 P'_c(\hat{g},x)) \tag{9}$$

where $-(P'_v(\hat{g},x)\log_2 P'_v(\hat{g},x) + P'_c(\hat{g},x)\log_2 P'_c(\hat{g},x))$ is the entropy computed only on void and cluster pixels, a close to 0 score indicating that the interpretable portion of the feature is mostly void or cluster. $\text{Purity}_{vc}(\hat{g},x)$ is 0 if among the pixels in the proposed feature that are either void or cluster pixels, half are void and half are cluster pixels, and 1 if all are void or all are cluster pixels, regardless of how many other pixels there are in the proposed feature.

We also have the ratio

$$\text{Ratio}_{vc}(\hat{g},x) = (P_v(\hat{g},x) + P_c(\hat{g},x)) \tag{10}$$

which is the total proportion of the feature that is any interpretable feature type at all.

We then have our ExpertAlign for Mass Maps:

$$\textsc{ExpertAlign}(\hat{g}, x) = \text{Purity}(\hat{g}, x) \cdot \text{Ratio}(\hat{g}, x) \tag{11}$$

which is then 0 when all the pixels in the feature are neither void or cluster, and 1 if all pixels are void pixels or all pixels are cluster pixels, and somewhere in the middle if most pixels are void or cluster pixels but there is a mix between both.

A.2 Supernova Dataset

Problem Setup. We extracted data from the PLAsTiCC Astronomical Classification challenge (Team et al., 2018). [2] PLAsTiCC dataset was designed to replicate a selection of observed objects with type information typically used to train a machine learning classifier. The challenge aims to categorize a realistic simulation of all LSST observations that are dimmer and more distorted than those in the training set. The dataset contains 15 classes, with 14 of them present in the training sample. The remaining class is intended to encompass intriguing objects that are theorized to exist but have not yet been observed.

In our dataset, we split the original training set into 90/10 training/validation, and the original test set was uploaded unchanged. We made these sets balanced for each class. The class includes objects such as tidal disruption event (TDE), peculiar type Ia supernova (SNIax), type Ibc supernova (SNIbc), and kilonova (KN). The dataset contains four columns: observation times (modified Julian days, MJD), wavelength (filter), flux values, and flux error. Spectroscopy measures the flux with respect to wavelength, similar to using a prism to split light into different colors.

Due to the expected high volume of data from upcoming sky surveys, it is not possible to obtain spectroscopic observations for every object. However, these observations are crucial for us. Therefore, we use an approach to capture images of objects through different filters, where each filter selects light within a specific broad wavelength range. The supernova dataset includes 7 different wavelengths that are used. The flux values and errors are recorded at specific time intervals for each wavelength. These values are utilized to predict the class that this data should be classified into.

Metric. We use the following expert alignment metric to measure if a group of features is interpretable:

$$\textsc{ExpertAlign}(\hat{g}, x) = \max_{w \in W} \text{LinearConsistency}(\hat{g}, x_w) \tag{12}$$

where W is the set of unique wavelength, $\hat{g}$ is the feature group, and x_w is the subset of x within wavelength w. In the supernova setting, there are three parameters: ϵ, the parameter for how much standard deviation σ is allowed, window size λ and the step size τ. Therefore, we formulate the LinearConsistency function as follows:

$$\text{LinearConsistency}(\hat{g}, x_w) = p(\hat{g}, x_w) \cdot d(\hat{g}, x_w) \tag{13}$$

$p(\hat{g}, x_w)$ is the percentage of data points that display linear consistency, penalized by $d(\hat{g}, x_w)$, which is the percentage of time steps containing data points.

Let $\beta(x, y) = \arg\min_\beta (X^T\beta - y)^2$, where $X = [x \quad 1]$ and $\beta = [\beta_1 \quad \beta_0]$. Here, β_1 is the slope and β_0 is the intercept. M is the number of data points in x_w, and $\hat{y}_{w,i} = x_{w,i} \cdot \beta$. Then, we have

$$\text{p}(\hat{g}, x_w) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}[\hat{y}_{w,i} \in [y_{w,i} - \epsilon \cdot \omega_{w,i}, y_{w,i} + \epsilon \cdot \omega_{w,i}]] \tag{14}$$

Let $t_1, ..., t_N$ be time steps at step size intervals. Then $t_i = t_{start} + i * \tau$, and N is the number of time steps. We also have

$$\text{d}(\hat{g}, x_w) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\exists_i : x_{w,i} \in [t_i, t_i + \lambda]] \tag{15}$$

A higher $\textsc{ExpertAlign}(\hat{g}, x) \in [0, 1]$ value means the flux slope at each wavelength is consistently linear and there are not many time intervals without data.

[2] https://www.kaggle.com/c/PLAsTiCC-2018

A.3 Multilingual Politeness Dataset

Problem Setup. This politeness dataset from Havaldar et al. (2023b) is intended for politeness classification, and would likely be solved via a fine-tuned multilingual LLM. Namely, this would be a regression task, using a trained LLM to output the politeness level of a given conversation snippet as a real number ranging from -2 to 2.

The dataset is accompanied by a theory-grounded politeness lexica. Such lexica built with domain expert input have been promising for explaining style (Danescu-Niculescu-Mizil et al., 2013), culture (Havaldar et al., 2024), and other such complex multilingual constructs.

Metric. Assume a theory-grounded Lexica L with k categories: $L = \ell_1, \ell_2, ...\ell_k$, where each set $\ell_i \subseteq \mathcal{W}$, where $\mathcal{W}$ is the set of all words. For each category, we use an LLM to embed all the contained words and then average the resulting embeddings, to get a set C of k centroids: $C = c_1, c_2, ...c_k$. We define this formally as:

$$C : \left\{ \frac{1}{|\ell_i|} \sum_{w \in l_i} \text{embedding}(w) \text{ for all } i \in [1, k] \right\} \tag{16}$$

For a group $\hat{g}$ containing words $w_1, w_2, ...$, the group-level expert alignment score can be computed as follows:

$$\textsc{ExpertAlign}(\hat{g}, x) = \max_{c \in C} \frac{1}{|\hat{g}|} \sum_{w \in \hat{g}} \cos(\text{embedding}(w), c) \tag{17}$$

Note that each language has a different theory-grounded lexicon, so we calculate a unique domain alignment score for each language.

A.4 Emotion Dataset

Problem Setup. This dataset is intended for emotion classification and is currently solved with a fine-tuned LLM (Demszky et al., 2020). Namely, this is a classification task where an LLM is trained to select some subset of 28 emotions (including neutrality) given a 1-2 sentence Reddit comment.

Axis Anchor	**Russell Emotions**
Positive valence (PV)	Happy, Pleased, Delighted, Excited, Satisfied
Negative valence (NV)	Miserable, Frustrated, Sad, Depressed, Afraid
High arousal (HA)	Astonished, Alarmed, Angry, Afraid, Excited
Low arousal (LA)	Tired, Sleepy, Calm, Satisfied, Depressed

Table 3: Emotions used to define the valence and arousal axis anchors for projection into the Valence-Arousal plane. We select the 5 emotions from the circumplex closest to each axis point.

Projection onto the Circumplex. To define the valence and arousal axes, we first generate four axis-defining points by averaging the contextualized embeddings ("I feel [emotion]") of the emotions listed in Table 3. This gives us four vectors in embedding space – positive valence ($\vec{v}_{\text{pos}}$), negative valence($\vec{v}_{\text{neg}}$), high arousal($\vec{a}_{\text{high}}$), and low arousal($\vec{a}_{\text{low}}$). We mathematically describe our projection function below:

1. We define the valence axis, V, as $\vec{v}_{\text{pos}} - \vec{v}_{\text{neg}}$ and the arousal axis, A, as $\vec{a}_{\text{high}} - \vec{a}_{\text{low}}$. We then normalize V and A and calculate the origin as the midpoints of these axes: $(\vec{v}_{\text{middle}}, \vec{a}_{\text{middle}})$.
2. We then scale the axes so $\vec{v}_{\text{pos}}$, $\vec{v}_{\text{neg}}$, $\vec{a}_{\text{high}}$, and $\vec{a}_{\text{low}}$ anchor to $(1, 0)$, $(-1, 0)$, $(0, 1)$, and $(0, -1)$ respectively. This enforces the circumplex to be a unit circle in the valence-arousal plane.
3. We compute the angle θ between the valence-arousal axes by solving $\cos\theta = \frac{V \cdot A}{\|V\| \cdot \|A\|}$
4. For each embedding vector $\vec{x}$ in the set $\{x_i\}_{i=1}^n$ we want to project into our defined plane, we compute the valence and arousal components for x_i as follows:

$x_i^v = (x_i - \vec{v}_{\text{middle}}) \cdot \vec{V}$
$x_i^a = (x_i - \vec{a}_{\text{middle}}) \cdot \vec{A}$.

5. We calculate the x and y coordinates to plot, enforcing orthogonality between the axes:
$\tilde{x_i^v} = x_i^v - x_i^a \cdot \cos\theta$
$\tilde{x_i^a} = x_i^a - x_i^v \cdot \cos\theta$

6. Finally, we plot $(\tilde{x_i^v}, \tilde{x_i^v})$ in the Valence-Arousal plane. We then calculate the shortest distance from $(\tilde{x_i^v}, \tilde{x_i^v})$ to the circumplex unit circle.

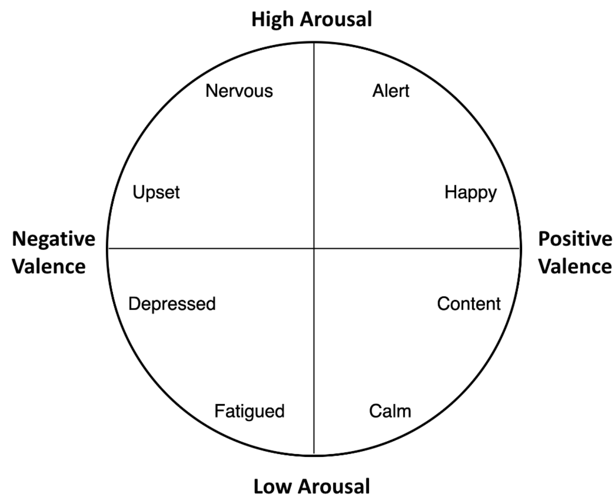


Figure 7: The circumplex model of affect Russell (1980).

Metric. We calculate the following two values for a proposed feature $\hat{g}$ containing words $w_1, w_2, ...$, where n is the number of words in $\hat{g}$:

$$\text{Signal}(\hat{g}) = \frac{1}{n} \sum_{w \in \hat{g}} |\|\text{Proj}(w)\|_2 - 1| \tag{18}$$

$$\text{Relatedness}(\hat{g}) = \frac{1}{n^2} \sum_{i}^{n} \sum_{j}^{n} \|\text{Proj}(w_i) - \text{Proj}(w_j)\|_2 \tag{19}$$

where $\text{Signal}(\hat{g}, x)$ measures the average Euclidean distance to the circumplex for every projected feature in $\hat{g}$, and $\text{Relatedness}(\hat{g}, x)$ measures the average pairwise distance between every projected feature in $\hat{g}$. We formalize the expert alignment metric as follows. For a group $\hat{g}$, the expert alignment score can be computed by:

$$\textsc{ExpertAlign}(\hat{g}, x) = \tanh(\exp[-\text{Signal}(\hat{g}, x) \cdot \text{Relatedness}(\hat{g}, x)]) \tag{20}$$

A.5 Chest X-Ray Dataset

We used datasets and pretrained models from TorchXRayVision (Cohen et al., 2022).[3] In particular, we use the NIH-Google dataset (Majkowska et al., 2020), which is a relabeling of the NIH ChestX-ray14 dataset (Wang et al., 2017). This dataset contains 28,868 chest X-ray images labeled for 14 common pathology categories, with a train/test split of 23,094 and 5,774. We additionally used a pre-trained structure segmentation model to produce 14 segmentations. The task is a multi-label classification problem for identifying the presence of each pathology. The 14 pathologies are:

> Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, Pneumothorax

[3] `https://github.com/mlmed/torchxrayvision`

The 14 anatomical structures are:

> Left Clavicle, Right Clavicle, Left Scapula, Right Scapula, Left Lung, Right Lung, Left Hilus Pulmonis, Right Hilus Pulmonis, Heart, Aorta, Facies Diaphragmatica, Mediastinum, Weasand, Spine

A.6 Laparoscopic Cholecystectomy Surgery Dataset

We use the open-source subset of the data from (Madani et al., 2022), which consists of surgeon-annotated video data taken from the M2CAI16 workflow challenge (Stauder et al., 2016) and Cholec80 (Twinanda et al., 2016) datasets. The task is to identify the safe/unsafe regions of where to operate. Specifically, each pixel of the image has one of three labels: background, safe, or unsafe. The expert labels provide each pixel with one of four labels: background, liver, gallbladder, and hepatocystic triangle.

B Interpretable Feature Extraction Details

Figure 8 illustrates a graphical model representing the Interpretable Feature Extraction pipeline for a given FIX dataset.

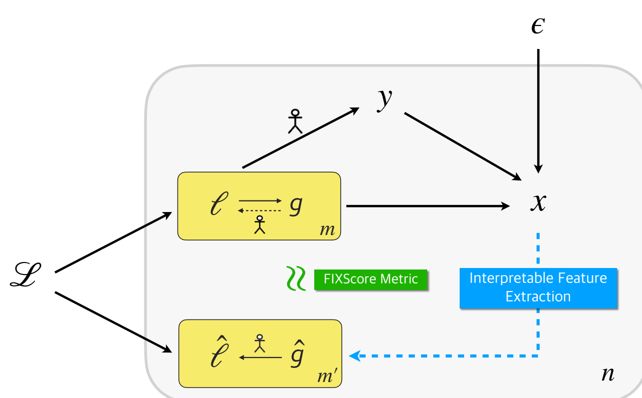


Figure 8: We illustrate a graphical model representing the Interpretable Feature Extraction pipeline for a given FIX dataset, with FIXScore metric in its general form. There are m true feature groups g and m latent features ℓ, and m' proposed feature groups $\hat{g}$ and m' proposed latent features $\hat{\ell}$. m does not have to equal m'. Moreover, n indicates the number of examples in the dataset. The person figure on near the closest arrow indicates that a domain expert would be able to infer the variable on the right-hand side of the arrow from the variable on the left-hand side arrow. In addition, ϵ is included to account for noise.

C Baselines Details

The FIX benchmark is publicly available at: `https://anonymous_website.com`.

Bootstrapping. For each setting's baselines experiments, we use a bootstrapping method (with replacement) to estimate the standard deviation of the sample means of FIXScore.

Group Maximum. For the number of groups, we take the scaling factor multiplied by the size of the distinct expert feature, which differs for each setting. The scaling factor we choose across all setting is 1.5 (and round up to the next nice whole number).

In the case of a supernova setting, we consider a distinct expert feature size of 6. This is because the maximum number of distinct expert features we can obtain is 6, given that there are a maximum of

	Method	Cholecystectomy	Chest X-ray	Mass Maps
Image	Identity	0.4686 ± 0.0096	0.2154 ± 0.0027	0.5486 ± 0.0033
	Random	0.1086 ± 0.0004	0.0427 ± 0.0001	0.5508 ± 0.0015
	Patch	0.0323 ± 0.0001	0.0999 ± 0.0008	0.5549 ± 0.0009
	Quickshift	0.2622 ± 0.0034	0.3419 ± 0.0025	0.5496 ± 0.0030
	Watershed	0.2807 ± 0.0051	0.1452 ± 0.0017	0.5594 ± 0.0011
	SAM	0.3678 ± 0.0074	0.3151 ± 0.0064	0.5526 ± 0.0009
	CRAFT	0.0271 ± 0.0007	0.1175 ± 0.0011	0.3991 ± 0.0017
Domain-Agnostic	Clustering	0.2880 ± 0.0049	0.2627 ± 0.0039	0.5518 ± 0.0009
	Archipelago	0.3351 ± 0.0034	0.2148 ± 0.0009	0.5509 ± 0.0015
		Supernova		
Time Series	Identity	0.0152 ± 0.0011		
	Random	0.0358 ± 0.0021		
	Slice 5	0.0337 ± 0.0015		
	Slice 10	0.0555 ± 0.0044		
	Slice 15	0.0554 ± 0.0032		
Domain-Agnostic	Clustering	0.2622 ± 0.0037		
	Archipelago	0.2574 ± 0.0082		
		Multilingual Politeness	**Emotion**	
Text	Identity	0.6070 ± 0.0015	0.0103 ± 0.0001	
	Random	0.6478 ± 0.0012	0.0303 ± 0.0004	
	Words	0.6851 ± 0.0010	0.1182 ± 0.0003	
	Phrases	0.6351 ± 0.0010	0.0198 ± 0.0003	
	Sentences	0.6109 ± 0.0006	0.0120 ± 0.0002	
Domain-Agnostic	Clustering	0.6680 ± 0.0048	0.0912 ± 0.0005	
	Archipelago	0.6773 ± 0.0006	0.0527 ± 0.0008	

Table 4: Baselines of different FIX settings. We report the mean FIXSCORE for all examples in each setting, with standard deviations.

3 humps in the time series dataset. For each hump, there are both peaks and troughs, leading to a potential maximum of 6 distinct expert features.

For the multilingual politeness setting, the group maximum would be 40, which is the total number of lexical categories, 26, with the scaling factor multiplied in to give some flexibility.

For the emotion setting, the group maximum would be , which is the total number of lexical categories, 26, with the scaling factor multiplied in to give some flexibility.

For mass maps, the group maximum would be 25. We compute the maximum number of local maximums 7 on mass maps blurred with $\sigma = 3$ and local minimums 7 on mass maps blurred with $\sigma = 5$, which sums up to be 14. We can then multiply with the scaling factor to give some flexibility and then we round up to 25.

Baseline Parameters. For mass maps, we use the following parameters for baselines. For patch, we use 8×8 grid. For QuickShift, we use kernel size 5, max dist 10, and sigma 0.2. For watershed, we use min dist 10, compactness 0. For SAM, we use 'vit_h'. For Archipelago, we use the same Quickshift parameters for the Quickshift segmenter.

Baseline Results. We report the full baseline results with standard deviations in Table 4.

D Compute Resources

All experiments were conducted on two server machines, each with 8 NVIDIA A100 GPUs and 8 NVIDIA A6000 GPUs, respectively.

E SAFEGUARDS

The datasets and models that we use in this work are not high risk and are previously open-source and publicly available. In particular, for our medical settings which would pose the most potential safety concern, the datasets we sourced our FIX datasets from are already open-source and consists of de-anonymized images.

F DATASHEETS

We follow the documentation framework provided by Gebru et al. (2021) to create datasheets for the FIX datasets. We address each section per dataset.

F.1 MOTIVATION

For what purpose was the dataset created?

- **Mass Maps**: The original dataset, CosmoGridV1 (Kacprzak et al., 2023), was created to help with predicting the initial states of the universe in cosmology.
- **Supernova**: The original dataset PLAsTiCC for Kaggle competition (Allam Jr et al., 2018), was created to classify astronomical sources that vary with time into different classes.
- **Multilingual Politeness**: The Multilingual Politeness dataset (Havaldar et al., 2023a) was created to holistically explore how politeness varies across different languages.
- **Emotion**: The original dataset, GoEmotions (Demszky et al., 2020), was created to help understand emotion expressed in language.
- **Chest X-Ray**: The NIH-Google dataset (Majkowska et al., 2020), which is a relabeling of the NIH ChestX-ray14 dataset (Wang et al., 2017), was created to help identify the presence of common pathologies.
- **Laparoscopic Cholecystectomy Surgery**: The original datasets from M2CAI16 workflow challenge (Stauder et al., 2016) and Cholec80 (Twinanda et al., 2016) were created to help identify the safe and unsafe areas of surgery.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

- **Mass Maps**: The original dataset CosmoGridV1 (Kacprzak et al., 2023) was created by Janis Fluri, Tomasz Kacprzak, Aurel Schneider, Alexandre Refregier, and Joachim Stadel at the ETH Zurich and the University of Zurich. The simulations were run at the Swiss Supercomputing Center (CSCS) as part of the project "Measuring Dark Energy with Deep Learning", hosted at ETH Zurich by the IT Services Group of the Department of Physics. We adapt the dataset and add a validation split.
- **Supernova**: The original dataset PLAsTiCC was created by Team et al. (2018). We adapt the dataset, add a validation split, and balance the sets for each class.
- **Multilingual Politeness**: The Multilingual Politeness dataset (Havaldar et al., 2023a) was created by Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar at the University of Pennsylvania.
- **Emotion**: The original GoEmotions (Demszky et al., 2020) dataset was created by Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi at Stanford University, Google Research and Amazon Alexa.
- **Chest X-Ray**: The NIH-Google dataset (Majkowska et al., 2020) was created by Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al., at Google Health, Stanford Healthcare and Palo Alto Veterans Affairs, Apollo Radiology International, and California Advanced Imaging.
- **Laparoscopic Cholecystectomy Surgery**: The M2CA116 workflow challenge dataset (Stauder et al., 2016) was created by Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feußner, and Nassir Navab at Technische Universität München in Germany and Johns Hopkins University. The Cholec80 dataset (Twinanda et al., 2016) was created by Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas

Padoy, at ICube, University of Strasbourg, CNRS, IHU, University Hospital of Strasbourg, IRCAD and IHU Strasbourg, France.

Who funded the creation of the dataset?

- Please refer to each setting's respective papers for funding details.

F.2 Composition

- The answers are described in our paper. Please refer to Section 4 and Appendix A for more details.

F.3 Collection Process

- We defer the collection process to the relevant works that created them. Please refer to Section 4 and Appendix A for more details.

F.4 Preprocessing/cleaning/labeling

- The answers are described in our paper. Please refer to Section 4 and Appendix A for more details.

F.5 Uses

- The answers are described in our paper. Please refer to Section 4 and Appendix A for more details.

F.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

- No. Our datasets will be managed and maintained by our research group.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

- The FIX datasets are released to the public and hosted on Huggingface (please refer to links in Appendix A).

When will the dataset be distributed?

- The datasets have been released now, in 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

- **Mass Maps**: The Mass Maps dataset is distributed under CC BY 4.0, following the original dataset CosmoGridV1 (Kacprzak et al., 2023).
- **Supernova**: The Supernova dataset is distributed under the MIT license.
- **Multilingual Politeness**: The Multilingual Politeness dataset is distributed under the CC-BY-NC license.
- **Emotion**: The Emotion dataset is distributed under the Apache 2.0 license.
- **Chest X-Ray**: The Chest X-Ray dataset is distributed under the Apache 2.0 license.
- **Laparoscopic Cholecystectomy Surgery**: The Laparoscopic Cholecystectomy Surgery dataset is distributed under the CC by NC SA 4.0 license.

G Author Statement

We bear all responsibility for any potential violation of rights, etc., and confirmation of data licenses.