NJUAI 南京大学 人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY

LAMDA
Learning And Mining from DatA

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20-27, 2026 | SINGAPORE

Distribution-Based Feature Attribution for Explaining the Predictions of Any Classifier

Xinpeng **Li**, **K**ai **M**ing **Ting**

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

School of Artificial Intelligence, Nanjing University, Nanjing, China

lixp@lamda.nju.edu.cn, tingkm@nju.edu.cn

The Current State of Feature Attribution

The Challenge:

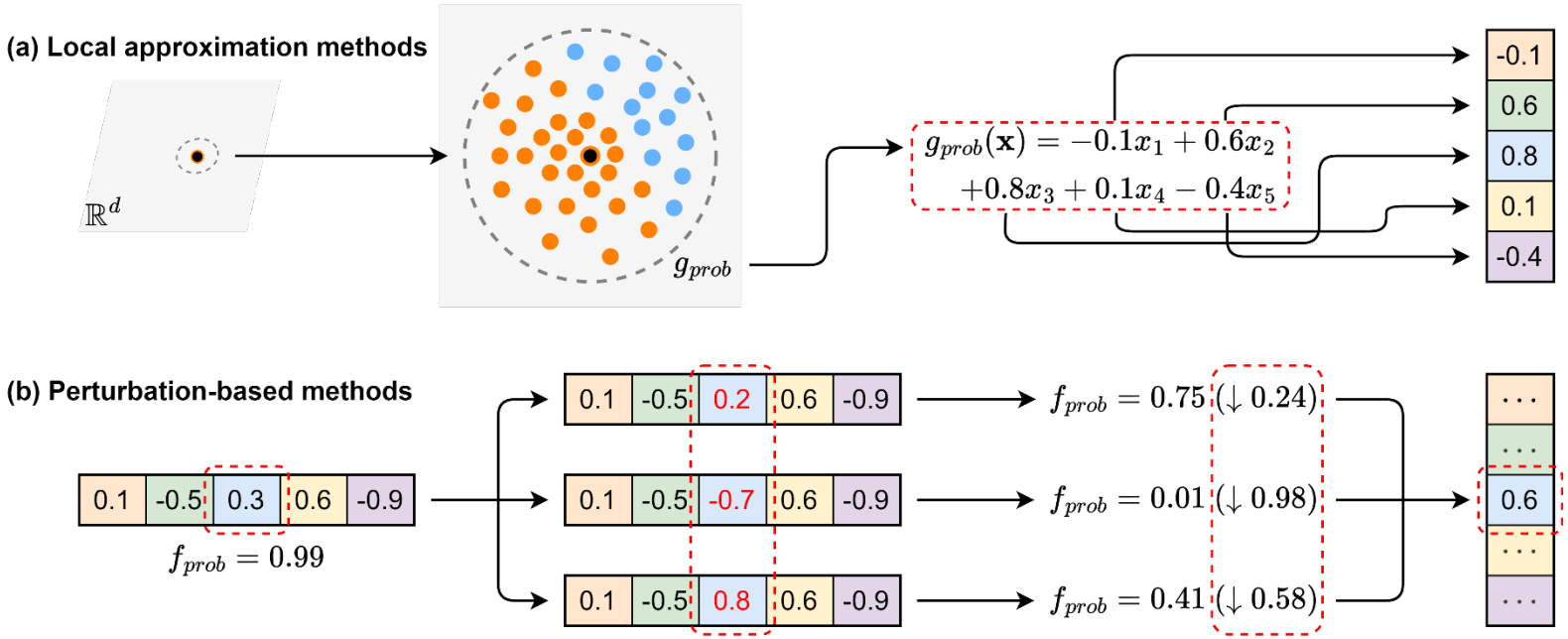
- Complex “black-box” AI models are popular but opaque.
- Feature Attribution is the dominant post-hoc solution (*e.g.*, SHAP, LIME).

The Gap:

- Despite popularity, the field has historically lacked a formal problem definition.
- Consequence: Methods are designed intuitively but often violate the underlying data distribution.

Why Existing Methods Fail

Many methods create synthetic instances to probe the model.



Building the explanatory model using OOD data produces explanations based on a distribution where the model's behavior is irrelevant or inapplicable, which invalidates the goal of understanding the model's behavior on its operational distribution.

Problem Definition

Assume that the target instance $\mathbf{x}^*$, the given dataset $\mathbf{X}$, and the training set $\mathbf{D}$ are all i.i.d. samples from the same underlying probability distribution $\mathcal{P}$. When the objective is to **understand the model's behavior on its operational data distribution**, the problem of feature attribution can be defined as:

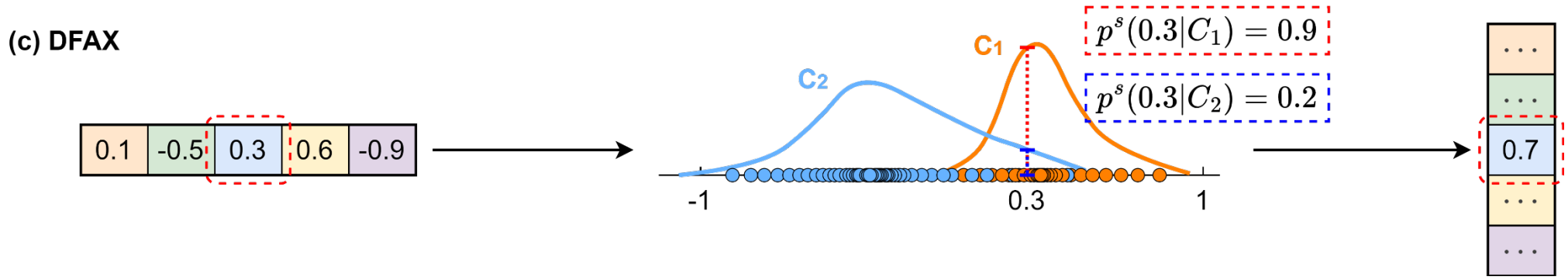
Definition 1: Feature Attribution

For a target instance $\mathbf{x}^* \sim \mathcal{P}$ with features $\mathcal{A}$ whose prediction $y^* = f(\mathbf{x}^*)$ is produced by classifier f, the task of feature attribution aims to provide an explanation as a score $I(\mathbf{x}^*, s|\mathbf{X})$ to each feature $s \in \mathcal{A}$. This score quantifies the influence of the specific feature-value, $\mathbf{x}^*_s$, on the classifier f to produce the prediction y^*, where a higher score indicates a greater influence towards this prediction. The explanatory model, $I(\cdot|\mathbf{X})$, must be built directly from the dataset $\mathbf{X}$, which reflects the underlying distribution $\mathcal{P}$, and the score $I(\mathbf{x}^*, s|\mathbf{X})$ is valid if and only if it is supported by $\mathcal{P}$.

Problem Definition

Key Criterion: The explanatory model and its explanation must be supported by the distribution $\mathcal{P}$ which is represented by the unmodified dataset X. Any modification to X that changes the underlying distribution, or the use of OOD instances, invalidates the feature attribution.

Proposed Method: DFAX



Definition 2: Distributional Feature Attribution eXplanations (DFAX)

Given the target instance $\mathbf{x}^*$ and feature $s \in \mathcal{A}$, DFAX computes the score as the difference between the conditional probability of $\mathbf{x}^*$ given the target class and that given all the other classes:

$$I(\mathbf{x}^*, s|\mathbf{X}) = p^s(\mathbf{x}^* \mid \{y^*\}) - p^s(\mathbf{x}^* \mid [m] \setminus \{y^*\}) = K^s\left(\mathbf{x}^*|\mathbf{X}_{\{y^*\}}\right) - K^s\left(\mathbf{x}^*|\mathbf{X} \setminus \mathbf{X}_{\{y^*\}}\right)$$

where the probability is computed using a kernel density estimator (KDE) K^s in the 1D subspace defined by feature s.

- **Logic:** How much does this feature value characterize the **target class** vs. **all other classes**?

Characteristics of DFAX

1. **Distributionally Supported:** DFAX complies with the problem definition. Explanations are always supported by the underlying data distribution $\mathcal{P}$ as the unmodified $\mathbf{X}$ is used to estimate conditional probability.
2. **Model-Agnostic & Decoupled:** DFAX operates solely on $\mathbf{X}$ and its pre-computed predictions, eliminating the need for further queries to the black-box classifier.
3. **Global Context:** DFAX makes full use of the global information contained in the entire dataset $\mathbf{X}$, rather than limiting itself to a local region that corresponds only to a subset of $\mathbf{X}$.
4. **Computational Efficiency:** If the kernel used for density estimation has a finite-dimensional feature map, the kernel mean map of $\mathbf{X}$ can be pre-computed before the target instance is provided. This one-off pre-computation significantly accelerates the attribution process.

Quantitative Evaluation

Setup: 10 datasets (tabular, text, image) with different classifiers, 2 implementations of DFAX with distinct KDEs, 6 model-agnostic baselines.
Metrics: Deletion (↓ lower is better) & insertion (↑ higher is better) scores.

Result:

- DFAX (in both implementations) significantly outperforms other baselines, securing the best average scores and overall rankings.
- Ranks 1st or 2nd on 9 out of 10 datasets.

Method	Avg. Deletion (↓)	Del. Rank	Avg. Insertion (↑)	Ins. Rank
$DFAX_G$	**0.3244**	**1.5**	**0.7708**	**1.2**
$DFAX_S$	0.3344	1.6	0.7470	2.0
DLIME	0.4595	4.1	0.6612	3.8
LINEX	0.5287	4.8	0.6326	4.1
SLISE	0.5457	5.5	0.6068	5.1
SHAP	0.5246	5.4	0.5463	7.3
MAPLE	0.5671	6.1	0.5800	6.1
Random	0.5709	7.0	0.5838	6.4

Qualitative Evaluation

Rotten Tomatoes: Identify the most important word in each movie review snippet signifying its sentiment.

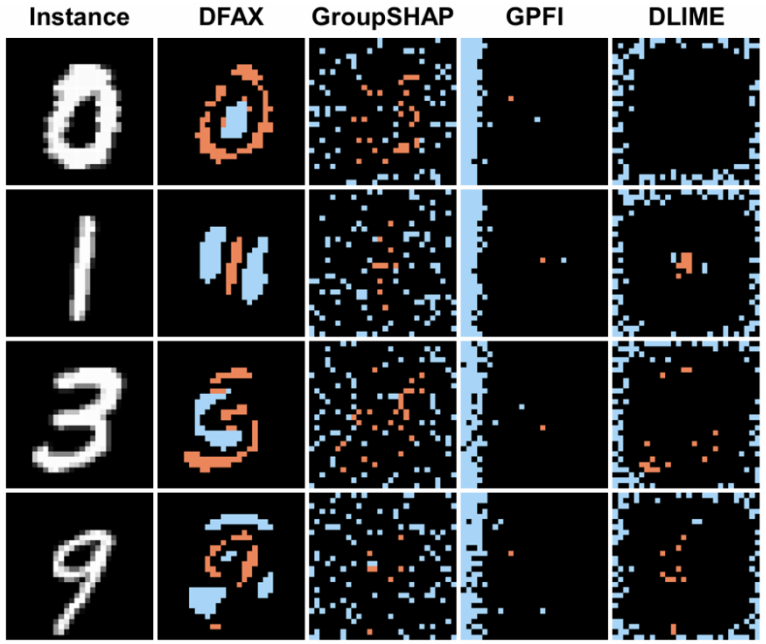
Positive example 1	Positive example 2	Positive example 3
the film's *real* appeal won't be to clooney fans or adventure buffs, but to moviegoers who enjoy thinking about **compelling** questions with no easy answers	a **fascinating**, bombshell *documentary* that should shame americans. . .	. . . **moving** film that respects its *audience* and its source material
Negative example 1	**Negative example 2**	**Negative example 3**
contains the *humor*, characterization, poignancy, and intelligence of a **bad** sitcom	a dark, **dull** *thriller* with a parting shot that misfires	. . . one resurrection **too** *many*

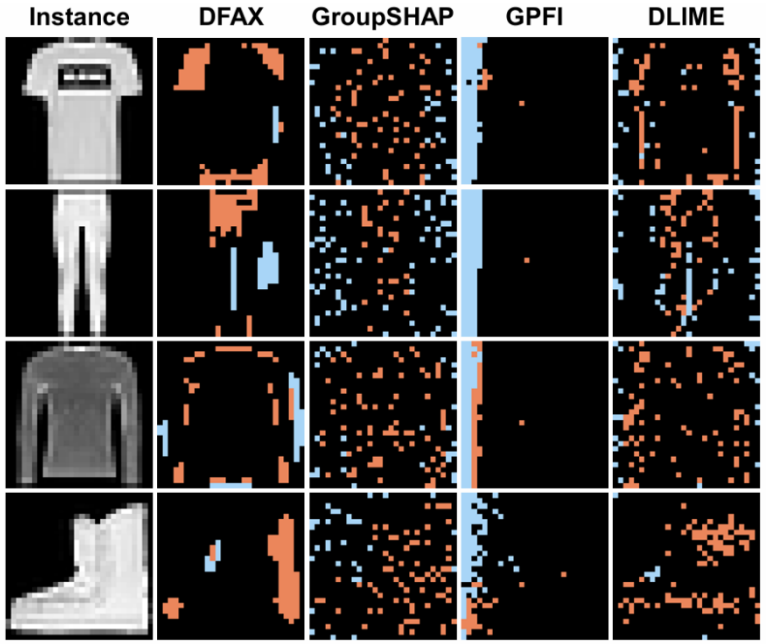
Table 3: Example snippets with positive and negative sentiments from the RottenTomatoes dataset. The most important word in a snippet found by **DFAX** is shown in boldface, while that found by *DLIME* is in italic.

DFAX identifies human-aligned, sentiment-carrying words.

Qualitative Evaluation

MNIST and FMNIST: For each image, identify the 100 most salient pixels based on their attribution scores.





- Pixels overlapping with the original image's foreground are shown in coral, while those in the background regions are colored light blue.
- DFAX selectively identifies semantically meaningful pixels crucial for the prediction, rather than simply selecting all non-zero pixels.

Qualitative Evaluation

Spatial Transcriptomics: For each cell, identify the top 50% of genes most critical for determining its tissue type. The non-salient genes are masked to verify if the classifier's original prediction is preserved.

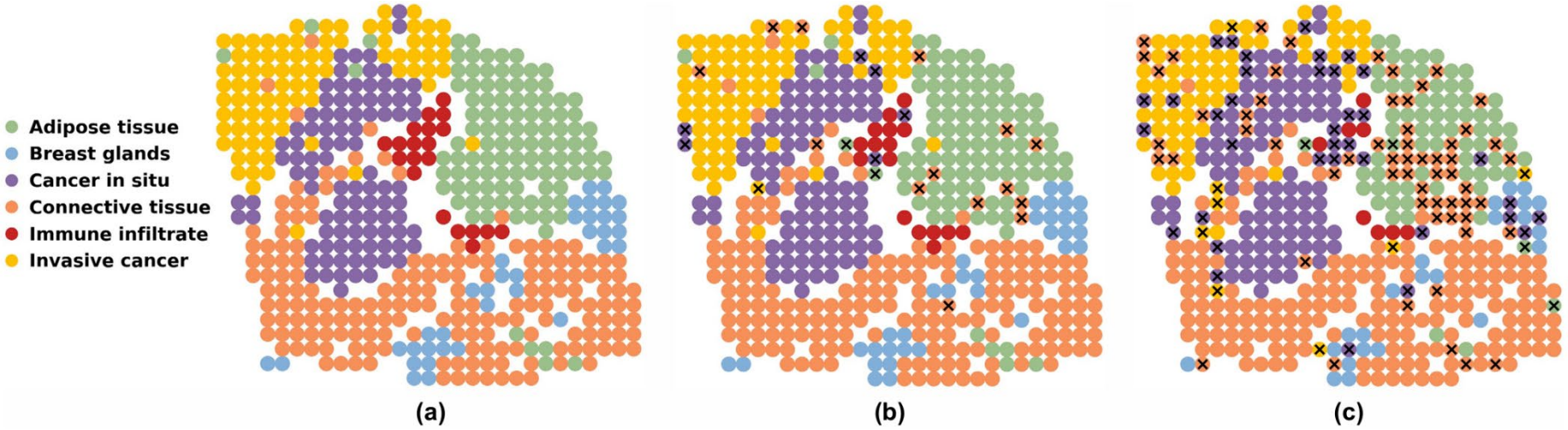
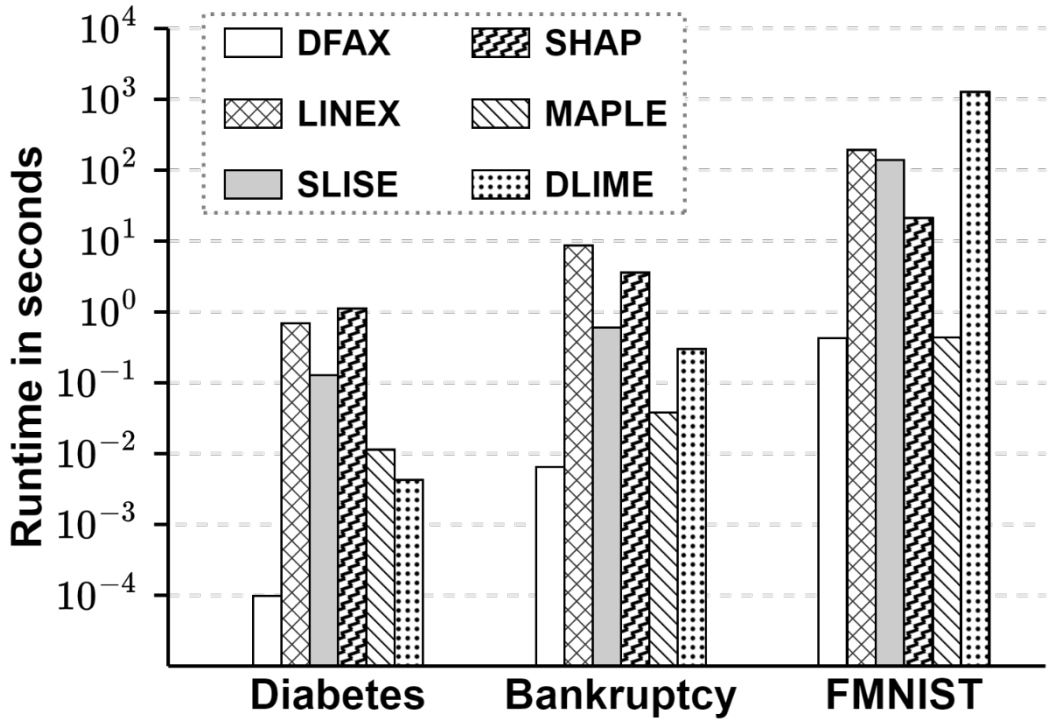


Figure 2: Visualization of tissue predictions on the HER2st dataset. (a) Initial predictions for the original cells with all 314 genes. (b-c) Predictions based on the 157 salient genes only, identified by (b) DFAX and (c) DLIME.

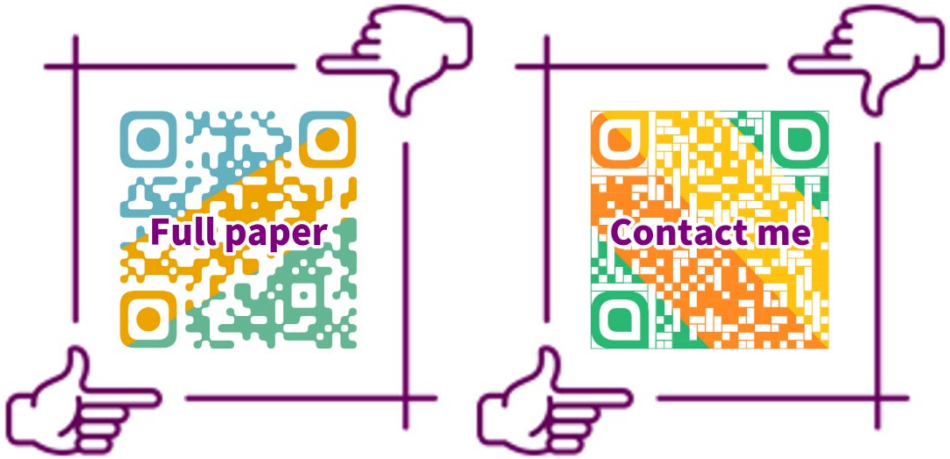
- DFAX achieves a high accuracy of 95.64%, significantly outperforming DLIME's 79.51%.
- Notably, DFAX results in only 6 misclassifications between normal and cancer cells, compared to 44 for DLIME.

Runtime Efficiency



- DFAX is often faster than other baselines by orders of magnitude, affirming its efficiency.

Thanks!



Presented by **Xinpeng Li**