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

**X**inpeng **Li**, **K**ai **M**ing **Ting**

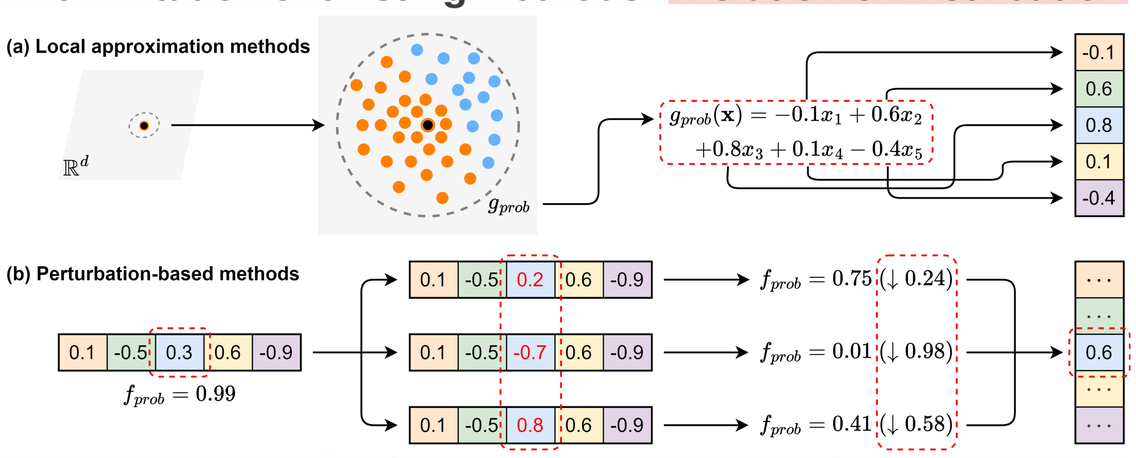
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
School of Artificial Intelligence, Nanjing University, Nanjing, China

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

## Motivation & Analysis

The proliferation of complex, black-box AI models has intensified the need for feature attribution methods capable of explaining their decisions. However, the field has historically lacked a formal problem definition.

**The limitation of existing methods: Violation of Distribution**




**Consequence: Explanations are often derived from data points belonging to a distribution where the model's behavior is irrelevant or inapplicable.**

## Problem Definition

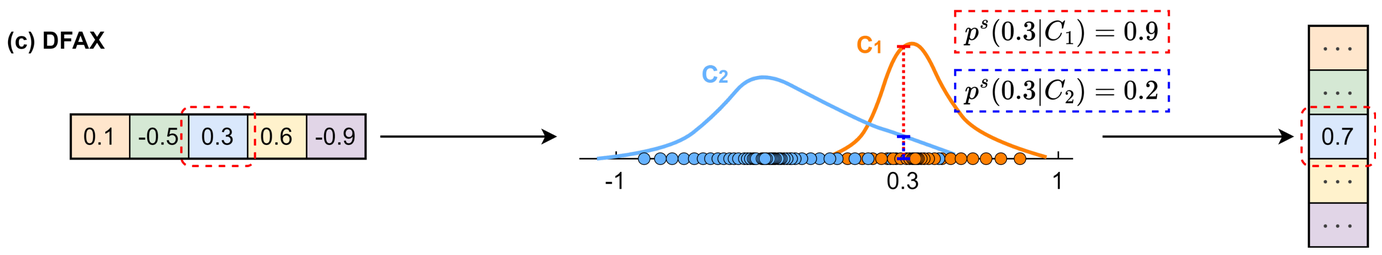
We assume that the target instance $\mathbf{x}^*$, the given dataset $\mathbf{X}$, and the training set $\mathbf{D}$ are all i.i.d. samples from the same underlying probability distribution $\mathcal{P}$. When the objective is to understand the model's logic on its operational data distribution, the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

### 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 $\mathcal{P}$.

**Key Criterion: The explanatory model and its explanation must be supported by the distribution $\mathcal{P}$ which is represented by the unmodified dataset $\mathbf{X}$.**

## Proposed Method: Distributional Feature Attribution eXplanations (DFAX)




- **Distributionally Supported**: Uses the unmodified $\mathbf{X}$ from the underlying distribution $\mathcal{P}$ to estimate conditional probability.
- **Model-Agnostic & Decoupled**: Operates solely on $\mathbf{X}$ and predictions; no need to query the black-box model repeatedly.
- **Global Context**: Makes full use of the entire dataset, rather than focusing on a local region.
- **Efficiency**: The kernel mean map of $\mathbf{X}$ can be one-off pre-computed to significantly speed up the attribution process.

### Definition 2: 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 $K^s$ in the one-dimensional subspace defined by feature $s$.

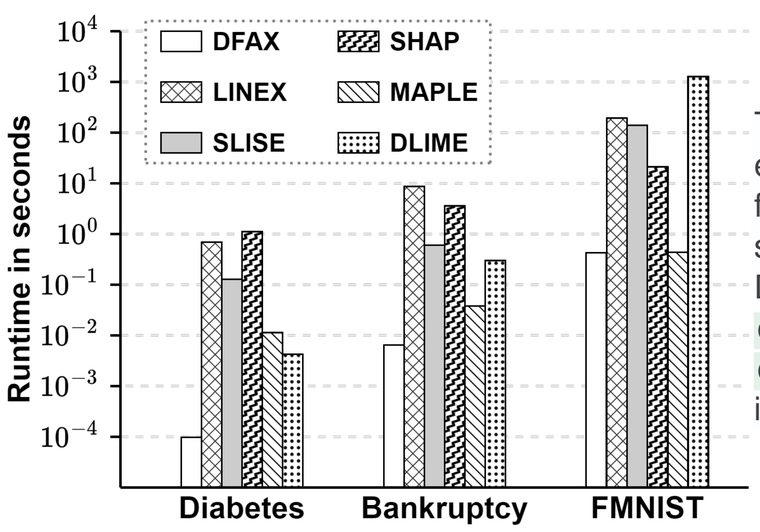
## Quantitative Evaluation

We compared two implementations of DFAX against 5 model-agnostic baselines on 10 diverse datasets (tabular, text, image) with various classifiers.

| 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 |
| MAPLE | 0.5671 | 6.1 | 0.5800 | 6.1 |
| SHAP | 0.5246 | 5.4 | 0.5463 | 7.3 |

**Result:** The proposed DFAX (in both implementations) consistently outperforms all baselines, securing the best average scores and overall rankings.

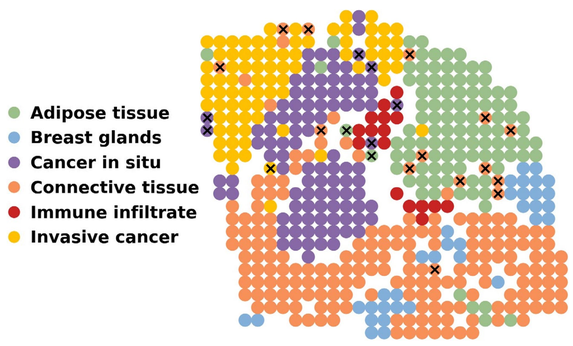
## Runtime Comparison




The total time required for each method to generate feature attributions for a single target instance. DFAX is often faster than other baselines by orders of magnitude, affirming its efficiency.

## Qualitative Evaluation

The 100 most salient pixels for MNIST and FMNIST images as identified by DFAX. Pixels overlapping with the original image's foreground are shown in coral, while those in the background regions are colored light blue. These results demonstrate that DFAX selectively identifies semantically meaningful pixels crucial for the prediction, rather than simply selecting all non-zero pixels.




Tissue type prediction preservation on spatial transcriptomics data, utilizing only the top 50% most important genes identified by the explainer. DFAX achieves a high accuracy of 95.64% in this masked evaluation.

## Conclusion

The proposed **DFAX** is the first fast and effective explainer to approach feature attribution by directly leveraging the underlying data distribution. This is made possible via our formal **problem definition** that addresses a foundational gap in the field.

Full paper

Contact me