

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

Xinpeng Li, Kai Ming 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

Abstract

The proliferation of complex, black-box AI models has intensified the need for techniques that can explain their decisions. Feature attribution methods have become a popular solution for providing post-hoc explanations, yet the field has historically lacked a formal problem definition. This paper addresses this gap by introducing a formal definition for the problem of feature attribution, which stipulates that explanations be supported by an underlying probability distribution represented by the given dataset. Our analysis reveals that many existing model-agnostic methods fail to meet this criterion, while even those that do often possess other limitations. To overcome these challenges, we propose Distributional Feature Attribution eXplanations (DFAX), a novel, model-agnostic method for feature attribution. DFAX is the first feature attribution method to explain classifier predictions directly based on the data distribution. We show through extensive experiments that DFAX is more effective and efficient than state-of-the-art baselines.

Extended version — <https://arxiv.org/abs/2511.09332>

Introduction

Recent years have witnessed a rapid growth and widespread adoption of artificial intelligence (AI). However, a major challenge remains: most state-of-the-art models are black-boxes which are unable to explain their own decisions. This lack of transparency has motivated the development of explainable AI (XAI) techniques, a field dedicated to helping users understand and trust these models (Minh et al. 2022).

Feature attribution has emerged as a crucial XAI technique for providing post-hoc explainability by computing contribution scores, which quantify the importance of input features with respect to the output of a model (Arrieta et al. 2020). These methods are broadly categorized as either model-specific or model-agnostic. The model-specific approach leverages the knowledge concerning the model’s internal structure. For example, numerous methods have been proposed for deep neural networks (DNNs) (Zhu et al. 2024; Walker et al. 2024), and they often require access to differentiable gradients in DNN. This reliance restricts their applicability to DNN only. In contrast, model-agnostic methods

treat a model as a black-box, generating feature attributions solely by observing and analyzing the model’s input-output behavior. This approach makes them universally applicable to any model, regardless of model types. Major families of model-agnostic feature attribution methods include local approximation and perturbation-based approaches (Li et al. 2023; Ivanovs, Kadikis, and Ozols 2021).

Despite years of research and the development of numerous methods, the field of feature attribution has, until now, lacked a formal problem definition. This foundational gap complicates the analysis and comparison of different methods. In this paper, we are motivated to provide a formal definition for the task of feature attribution. Our proposed definition serves as a foundational criterion, establishing a clear standard for the analysis of specific methods.

Furthermore, our analysis of existing methods reveals that even those that satisfy our proposed problem definition often suffer from other limitations restricting their performance or practical applicability. To address these issues, we introduce Distributional Feature Attribution eXplanations (DFAX), a novel model-agnostic method for explaining classifier predictions based on the probability distribution of a given dataset. DFAX is designed to adhere to our formal problem definition while simultaneously overcoming the key limitations of prior approaches.

The main contributions of this work are:

- We introduce a formal definition for the problem of feature attribution, which provides a key criterion for the analysis and design of feature attribution methods.
- With this problem definition, we conduct an analysis of existing methods, identifying their individual limitations.
- We propose the DFAX scheme which complies with the definition and does not have the identified limitations of prior methods. To the best of our knowledge, DFAX is the first explainer to approach feature attribution by directly leveraging the underlying data distribution.
- Through extensive quantitative and qualitative experiments, we demonstrate the superior effectiveness and efficiency of DFAX compared to other state-of-the-art model-agnostic methods for feature attribution.

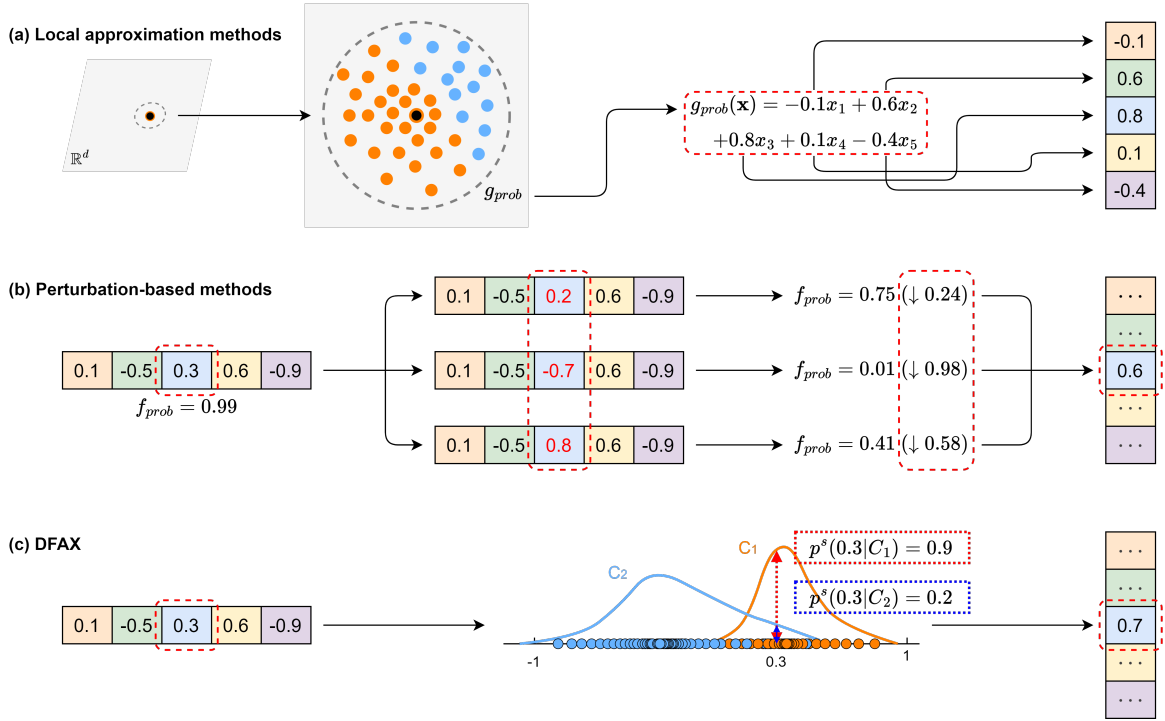


Figure 1: Illustrations of two existing families of methods, shown in (a) & (b); and the proposed DFAX, shown in (c).

Related Work

This section reviews the two predominant families of model-agnostic feature attribution methods, followed by an overview of kernel density estimation, a core component of our proposed DFAX method.

Local Approximation Methods

Local approximation methods, as illustrated in Figure 1(a), operate by first defining a neighborhood around the target instance and then fitting a simple surrogate model to the classifier’s predictions within this local region. Feature attributions are then derived from this local surrogate model.

A prominent example is LIME (Ribeiro, Singh, and Guestrin 2016), which generates its neighborhood by creating synthetic points around the target instance via random sampling before fitting an explanatory linear model. DLIME (Zafar and Khan 2021), a deterministic version of LIME, leverages agglomerative hierarchical clustering on the training data to define the neighborhood. Similarly, MAPLE (Plumb, Molitor, and Talwalkar 2018) and SLISE (Björklund et al. 2023) select neighboring points from the given dataset instead of generating synthetic points as done in LIME. MAPLE weights these neighbors to produce more faithful local explanations, while SLISE fits a sparse, locally linear regression model whose coefficients serve as feature attributions. Focusing on improving stability and unidirectionality, LINEX (Dhurandhar et al. 2023) minimizes sensitivity to perturbations in a way inspired by the invariant risk minimization (IRM) principle. The specific implementation of LINEX can vary, as the local environments for IRM may

be created differently depending on the application scenario.

Perturbation-Based Methods

As depicted in Figure 1(b), perturbation-based methods operate by perturbing a feature’s value and measuring the resulting degradation in the classifier’s performance (e.g., the decrease in the predicted probability for the target class).

A famous method in this category is SHAP (Lundberg and Lee 2017), which is grounded in cooperative game theory and calculates feature importance using Shapley values (Shapley 1953), the marginal contribution of each feature to the prediction for the target instance. The practical implementation of SHAP varies based on how this contribution is estimated. A theoretically pure approach, based on Shapley regression values (Lipovetsky and Conklin 2001), requires retraining the classifier on all possible subsets of features. Common implementations instead use Shapley sampling values (Strumbelj and Kononenko 2014). These methods avoid retraining by using the original classifier trained with all feature present, and observing its performance changes on perturbed instances created by combining feature-values from the target instance and a background dataset.

Another perturbation-based method is PFI (Fisher, Rudin, and Dominici 2019), a model-agnostic version of the original PFI (Breiman 2001). PFI measures the importance of a feature by the expected loss in classifier performance after permuting the feature with values across the entire dataset.

Kernel Density Estimation

A kernel density estimator (KDE) is a non-parametric method for estimating the density/probability of each point

in a given dataset (Rosenblatt 1956; Ting et al. 2021). Given a dataset $\mathbf{X} \subset \mathbb{R}^d$, the KDE at any point $\mathbf{x}^* \in \mathbb{R}^d$ is defined as:

$$K(\mathbf{x}^*|\mathbf{X}) = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} \kappa(\mathbf{x}^*, \mathbf{x})$$

where κ is a point kernel. Different choices of κ yield different estimators. For example, a Gaussian kernel results in the Gaussian Kernel Density Estimator (GKDE), while the Isolation Kernel (Ting, Zhu, and Zhou 2018) produces Isolation Kernel Density Estimator (IKDE) (Ting et al. 2021).

A fast alternative to GKDE in outlying aspects mining tasks (Wells and Ting 2019) is SiNNE (Samariya et al. 2020), which is a simplified version of iNNE (Bandaragoda et al. 2014). While originally designed to estimate anomaly scores, SiNNE can be used as an efficient substitute for KDE in our feature attribution task.

The KDE computation can be significantly accelerated if the point kernel can be approximated as $\kappa(x, y) \approx \langle \varphi(x), \varphi(y) \rangle$ via some technique like the Nyström method (Williams and Seeger 2000), where φ is a finite-dimensional feature map approximating the feature map of κ . With this, the KDE can be re-expressed as:

$$K(\mathbf{x}^*|\mathbf{X}) \approx \langle \varphi(\mathbf{x}^*), \hat{\Phi}(\mathbf{X}) \rangle$$

where $\hat{\Phi}(\mathbf{X}) = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} \varphi(\mathbf{x})$ is the kernel mean map of \mathbf{X} (Muandet et al. 2017). This allows any subsequent probability/density estimation to be performed in $\mathcal{O}(1)$ time after a one-off computation for the kernel mean map of \mathbf{X} .

Problem Definition

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:

Let $\mathcal{A} = \{s_j\}_{j=1}^d$ be the set of features and m be the total number of classes. Let f be a classifier trained on a dataset $\mathbf{D} \subset \mathbb{R}^d$, which maps inputs to either a predicted class or a predicted probability for a target class ($f: \mathbb{R}^d \mapsto [m]$ or $[0, 1]$). Let $\mathbf{x}^* \in \mathbb{R}^d$ be the target instance and $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ be a given dataset. We assume \mathbf{x}^* , \mathbf{X} , and \mathbf{D} are all independent and identically distributed (i.i.d.) samples from the same underlying probability distribution \mathcal{P} .

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} .*

A crucial tenet of this definition is the role of the dataset \mathbf{X} , which serves as an empirical representation of the underlying distribution \mathcal{P} . Any modification to \mathbf{X} that changes the underlying distribution, in the process of building the

explanatory model $I(\cdot|\mathbf{X})$, invalidates its feature attribution. This is because building the explanatory model using synthetic or out-of-distribution (OOD) instances produces explanations based on a distribution where the model’s behavior is irrelevant or inapplicable. In a nutshell, the key criterion of Definition 1 is that *the explanatory model $I(\cdot|\mathbf{X})$ and its explanation $I(\mathbf{x}^*, s|\mathbf{X})$ must be supported by the distribution \mathcal{P} which is represented by the unmodified dataset \mathbf{X} (no OOD instance is used for generating feature attributions).*

Analyses of Existing Methods

Definition 1 provides the criterion for assessing any feature attribution method. In this section, we analyze the two predominant families of model-agnostic methods.

(a) **Local approximation methods.** Some of these methods satisfy Definition 1. For example, given the training dataset $\mathbf{D} = \mathbf{X}$ for the classifier f , DLIME (Zafar and Khan 2021) selects a neighboring cluster around the target instance from \mathbf{X} , upon which it fits a linear regression model $LR(\mathbf{x}) = \omega\mathbf{x} + b$. LR works as the explanatory model I in Definition 1. For each feature $s_i \in \mathcal{A}$ of the target instance \mathbf{x}^* , the coefficient ω_i in LR corresponds to the score $I(\mathbf{x}^*, s_i|\mathbf{X})$ stated in the definition. Note that while DLIME utilizes the unmodified \mathbf{X} to build the explanatory model, thus satisfying Definition 1, it only uses a small subset of \mathbf{X} which corresponds to a local region that DLIME focuses on to build LR , due to its reliance on fitting a simple surrogate model in a selected local neighborhood. This is an inherent limitation of local approximation methods.

Other methods in this family, such as LIME (Ribeiro, Singh, and Guestrin 2016), not only share this limitation of a local focus but also fail to satisfy the criterion of Definition 1. This is because their explanatory models are fitted to a synthetic neighborhood generated via random perturbation, independent of distribution \mathcal{P} .

(b) **Perturbation-based methods.** Some of these methods, such as an implementation of SHAP (Lundberg and Lee 2017) that calculates the Shapley regression values (Lipovetsky and Conklin 2001), comply with Definition 1. The score it produces is given by:

$$I(\mathbf{x}^*, s|\mathbf{X}) = \sum_{S \subseteq \mathcal{A} \setminus \{s\}} M_S (F_{S \cup \{s\}}(\mathbf{x}^*) - F_S(\mathbf{x}^*)) \quad (1)$$

$$\text{where } M_S = \frac{|S|!(|\mathcal{A}| - |S| - 1)!}{|\mathcal{A}|!}$$

where $n!$ is the factorial of n , and F_S is the estimated probability function of the classifier f retrained using only a subset of features $S \subseteq \mathcal{A}$, with all other features withheld.

In this case, $\mathbf{D} = \mathbf{X}$ in the subspace defined by the feature subset S is used for building the explanatory model I , satisfying the criterion of being supported by \mathcal{P} in Definition 1.

However, calculating Shapley regression values is computationally infeasible. Practical implementations commonly use Shapley sampling values (Strumbelj and Kononenko 2014), which introduce two key approximations. First, instead of iterating through the entire power set of $\mathcal{A} \setminus \{s\}$ with 2^{d-1} feature subsets, Equation 1 is approximated by

sampling a tractable number of subsets. Second, to avoid repeatedly retraining the classifier, $F_S(\mathbf{x}^*)$ is replaced by $\mathbb{E} \left[f_{prob}(\tilde{\mathbf{x}}_S) \mid \tilde{\mathbf{x}}_S \in \tilde{\mathbf{X}}_S \right]$ which is the expected output of the classifier f_{prob} . As f_{prob} is trained on \mathbf{D} with all features, the dataset \mathbf{X} is modified for the feature subset S as follows: $\tilde{\mathbf{X}}_S$ contains instances originally the same as \mathbf{X} , except that for each instance $\tilde{\mathbf{x}}_S \in \tilde{\mathbf{X}}_S$, its feature-values in the subset S are replaced with those in the target instance \mathbf{x}^* , while the values for the remaining features are kept unchanged.

While enhancing efficiency, this approximation method of SHAP does not satisfy Definition 1, because it modifies \mathbf{X} into $\tilde{\mathbf{X}}_S$, mixing feature-values from \mathbf{x}^* and the instances in \mathbf{X} . These synthetic instances invalidate the key criterion of feature attribution stated in Definition 1, *i.e.*, they are not i.i.d. samples from the distribution \mathcal{P} for which the classifier was trained. Consequently, the explanation is inappropriate as it is not supported by \mathcal{P} . This is a common oversight for most methods in this family, a result of designing a method without knowing the problem definition.

Note that while methods like LIME and Shapley sampling values do not satisfy Definition 1, they address a different objective: understanding model behavior across the entire input space, rather than on its operational data distribution (Chen et al. 2020). However, this objective has limited practical applications. For most real-world XAI purposes, the primary goal is to understand model behavior specifically on the data distribution it was trained and qualified to operate on (Freiesleben and König 2023).

Definition 1 provides the criterion for assessing whether a method has any compliance issues. Our analysis above shows that while some existing methods satisfy the criterion stated in Definition 1, others do not. A serious oversight of existing methods is the misuse of \mathbf{X} , creating synthetic instances that violate the underlying distribution \mathcal{P} . This fundamentally conflicts with the objective of understanding a model’s logic on its operational data distribution. Another limitation of many methods is the inability to fully utilize \mathbf{X} . The above analyses reveal the importance of Definition 1, which serves as a guide in designing a reliable method that complies with the objective and makes the full use of \mathbf{X} .

Proposed Method: DFAX

With the above-mentioned limitations of existing methods, we are motivated to develop a better approach that not only solves the problem of feature attribution defined in Definition 1, but also overcomes the identified limitations. A simple yet effective solution is via a distributional approach.

Let the predicted class for the target instance be $y^* = f(\mathbf{x}^*)$, and the corresponding predictions for \mathbf{X} be $\{y_i\}_{i=1}^n = \{f(\mathbf{x}_i)\}_{i=1}^n$. In the subspace defined by any feature $s \in \mathcal{A}$, the distribution, *i.e.*, the conditional probability, for a subset of classes $\mathcal{C} \subseteq [m]$, denoted as $p^s(\cdot | \mathcal{C})$, can be estimated from the subset of points $\mathbf{X}_{\mathcal{C}} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{X} \text{ and } y_i \in \mathcal{C}\}$. Our proposed Distributional Feature Attribution eXplanations (DFAX) method is then defined as:

Definition 2 (DFAX). *Given the target instance \mathbf{x}^* and feature $s \in \mathcal{A}$, DFAX computes the score as the difference be-*

tween the conditional probability of \mathbf{x}^ given the target class and that given all the other classes:*

$$\begin{aligned} I(\mathbf{x}^*, s | \mathbf{X}) &= p^s(\mathbf{x}^* | \{y^*\}) - p^s(\mathbf{x}^* | [m] \setminus \{y^*\}) \\ &= K^s(\mathbf{x}^* | \mathbf{X}_{\{y^*\}}) - K^s(\mathbf{x}^* | \mathbf{X} \setminus \mathbf{X}_{\{y^*\}}) \end{aligned}$$

where the probability is computed using a KDE K^s in the one-dimensional subspace defined by feature s .

In contrast to the two major families of model-agnostic feature attribution methods, DFAX operates directly from a distributional perspective, as illustrated in Figure 1(c). By measuring the probabilities based on the feature s , conditional on some classes, DFAX quantifies the extent to which the feature-value of the target instance \mathbf{x}^* characterizes the data points belonging to the target class in \mathbf{X} , while it does not characterize points from all other classes at the same time. To the best of our knowledge, DFAX is the first feature attribution approach based on this principle.

DFAX satisfies Definition 1 as the unmodified \mathbf{X} from distribution \mathcal{P} is used to estimate conditional probability. The role of \mathbf{X} in DFAX is similar to that of the training set used in the k-nearest neighbors algorithm. Following the lazy learning approach, DFAX defers the computation of probability until the target instance is provided, rather than learning an explicit explanatory model beforehand.

Compared with existing methods that comply with Definition 1, 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} . Moreover, DFAX allows for significant computational acceleration in practice. If the kernel used for density estimation has (or can be approximated by) 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 dramatically speeds up the attribution process, especially when there are multiple target instances to explain.

DFAX has several other desirable characteristics. First, its feature attribution process is fully decoupled from the classifier. DFAX operates solely on \mathbf{X} and its pre-computed predictions, eliminating the need for further queries to the classifier f , which is ideal for scenarios where the classifier is expensive to query or unavailable due to privacy or proprietary concerns. By substituting predictions with ground-truth labels, this decoupling also enables DFAX to explain the data’s inherent class structure, independent of any specific classifier. Second, DFAX utilizes global distributional information, generating attributions based on characteristic properties of the target class and non-target classes. This is distinct from local methods that are often sensitive to neighborhood selection (Visani, Bagli, and Chesani 2020).

Experiments

In this section, we demonstrate the effectiveness and efficiency of the proposed DFAX method through both quantitative and qualitative evaluations.

Experimental Setup

All experiments were conducted on a server equipped with an AMD EPYC 7742 CPU, an NVIDIA RTX A6000 GPU,

Datasets	#inst	#feat	#cls	Classifier, acc
Diabetes	520	16	2	RF, .98
HER2st	527	314	6	RF, .73
Rice	3,810	7	2	LR, .91
Waveform	5,000	40	3	LR, .86
Bankruptcy	6,819	95	2	AB, .97
RottenTomatoes	10,662	300	2	NB, .76
Pendigits	10,992	16	10	SVM, .99
DryBean	13,611	16	7	SVM, .98
MNIST	70,000	784	10	MLP, .99
FMNIST	70,000	784	10	ResNet, .94

Table 1: Datasets and classifiers used in the experiments, along with the characteristics of the datasets and testing accuracies of the classifiers. #inst stands for the number of instances, feat for features, cls for classes, RF for Random Forest, LR for Logistic Regression, AB for AdaBoost, NB for Naive Bayes, SVM for Support Vector Machine, MLP for Multi-Layer Perceptron, and ResNet for Residual Network.

and 1TB of memory. The system was running Ubuntu 20.04.2 (Linux kernel 5.4.0) with Python 3.9.12. Unless specified otherwise, we employ GKDE as the default KDE in our implementation of DFAX.

Datasets. The experiments are conducted on ten real-world datasets covering diverse modalities including tabular, text, and image, many of which have been utilized in prior studies (Asuncion, Newman et al. 2007; LeCun et al. 2002; Xiao, Rasul, and Vollgraf 2017; Andersson et al. 2020; Pang and Lee 2005). All features within each dataset are standardized. We randomly select 100 samples from each dataset as the testing set, with the remaining samples forming the training set. The training set \mathbf{D} is used as \mathbf{X} when required by a feature attribution method, while the testing set also serves as the set of target instances whose attributions are sought in the quantitative experiments.

Classifiers. For each dataset, we train a classifier on the training set and use it for prediction of the target instances in the testing set. A wide range of classifiers, such as support vector machine with radial basis function kernel, ensemble methods, and neural networks, are employed in this process (Breiman 2001; Freund and Schapire 1995; He et al. 2016). Table 1 provides the details of the datasets and classifiers.

Baseline methods. We compare DFAX against five state-of-the-art model-agnostic feature attribution methods: LINEX (Dhurandhar et al. 2023), SLISE (Björklund et al. 2023), SHAP (Lundberg and Lee 2017), MAPLE (Plumb, Molitor, and Talwalkar 2018), and DLIME (Zafar and Khan 2021). These selected baselines represent the two predominant families of methods: local approximation and perturbation-based. SLISE, MAPLE, and DLIME comply with the criterion of Definition 1, while the implementations of LINEX and SHAP we use here do not. It should be noted that while variants of LINEX and SHAP exist that also satisfy this definition, they have other limitations that restrict their general use. Therefore, to ensure a practical compar-

ison, we evaluate the most widely applicable implementations of LINEX and SHAP. In addition, a random baseline is incorporated as a sanity check in the quantitative evaluation. The hyperparameter specifications for all methods and datasets are provided in Appendix A.

Quantitative Evaluation

First, we quantitatively verify the effectiveness of DFAX through extensive comparative experiments.

Evaluation metrics. We adopt the deletion and insertion scores (Petsiuk, Das, and Saenko 2018) as our evaluation metrics. The deletion score measures the Area Under the Curve (AUC) of the classifier’s predicted probability for the target class of a given instance, as important features found by a feature attribution method are progressively masked with random values drawn from a standard normal distribution, while the insertion score does so as features are incrementally reintroduced. These scores quantify the fidelity of the attribution to the classifier (Klein et al. 2024). They have been widely adopted in the literature as established metrics for evaluating feature attribution methods (Walker et al. 2024; Li et al. 2023; Zhu et al. 2024).

Results. Table 2 presents the quantitative evaluation results for different model-agnostic feature attribution methods based on the deletion and insertion scores, along with their average scores and average rankings across all datasets.

The proposed DFAX (DFAX_G and DFAX_S implementations) significantly outperforms other baseline methods, securing the first or second rank on nine of the ten datasets. The only exception is the Diabetes dataset, where the performances of the top three methods are comparable. Between the two implementations, DFAX_G generally shows better results than DFAX_S, especially in terms of the insertion score.

Among other baseline methods evaluated, DLIME exhibits the strongest performance, followed by LINEX. However, DFAX outperforms both of them by a large margin. Note that the average insertion scores for both SHAP and MAPLE fall below the random baseline, thus failing the sanity check. Some of the baseline methods, *i.e.*, LINEX and SHAP, do not comply with the criterion of being supported by the distribution \mathcal{P} stated in Definition 1. While others (DLIME, SLISE, and MAPLE) do, they do not fully exploit the information in \mathbf{X} .

Two key factors contribute to the outstanding performance of DFAX. First, DFAX derives feature attributions supported by the underlying data distribution \mathcal{P} which is empirically represented by the provided dataset \mathbf{X} . Second, DFAX makes full use of \mathbf{X} , leveraging global distributional information from the entire dataset.

Qualitative Evaluation

Here, we aim to show through several examples that DFAX produces more intuitive attributions with better quality.

RottenTomatoes. This dataset consists of movie review snippets from the Rotten Tomatoes website. After TF-IDF vectorizing the snippets and acquiring the trained classifier’s predictions, the task is to identify the most important word in

Metrics	Datasets	DFAX _G	DFAX _S	LINEX	SLISE	SHAP	MAPLE	DLIME	Random
Deletion Score ↓	Diabetes	.5442	.5420	.6918	.7058	.6634	.6744	<u>.5422</u>	.7182
	HER2st	.1247	<u>.2395</u>	.4393	.4358	.3911	.4725	<u>.4560</u>	.4672
	Rice	<u>.6387</u>	.6321	.6849	.6859	.7050	.8203	.7458	.7224
	Waveform	.2621	<u>.2650</u>	.4544	.4559	.4905	.5814	.4335	.6020
	Bankruptcy	.5944	<u>.6129</u>	.6166	.6164	.6765	.6169	.6277	.6312
	RottenTomatoes	<u>.1213</u>	.0682	.6120	.7395	.5484	.7397	.3138	.7034
	Pendigits	.2529	<u>.2645</u>	.3766	.4041	.3651	.3997	.2845	.4248
	DryBean	.3148	<u>.3263</u>	.4705	.4786	.5397	.4662	.4754	.4933
	MNIST	<u>.1608</u>	.1535	.5847	.6456	.4779	.5932	.4591	.6351
	FMNIST	.2299	<u>.2403</u>	.3565	.2894	.3887	.3069	.2567	.3112
	Average	.3244	<u>.3344</u>	.5287	.5457	.5246	.5671	.4595	.5709
Avg. Ranking		1.5	<u>1.6</u>	4.8	5.5	5.4	6.1	4.1	7.0
Insertion Score ↑	Diabetes	<u>.8738</u>	<u>.8727</u>	.7643	.7375	.7352	.7789	.8781	.7454
	HER2st	.8119	<u>.6962</u>	.5023	.5078	.4648	.4749	.4851	.4762
	Rice	.8028	.8136	.7809	.7800	.7260	.6459	.7142	.7389
	Waveform	.8906	<u>.8820</u>	.7415	.7338	.6371	.6364	.7697	.6154
	Bankruptcy	.6557	<u>.6387</u>	.6354	.6353	.5720	.6341	.6238	.6204
	RottenTomatoes	.9907	<u>.9904</u>	.8714	.7494	.6237	.7483	.9550	.7668
	Pendigits	.5903	<u>.5524</u>	.4462	.4171	.4024	.4146	.5407	.4095
	DryBean	.6681	<u>.6570</u>	.5246	.5236	.4425	.5323	.5371	.5039
	MNIST	.9269	<u>.9212</u>	.6293	.6622	.4794	.6137	.7128	.6494
	FMNIST	.4975	<u>.4454</u>	.4302	.3166	.3797	.3205	.3952	.3119
	Average	.7708	<u>.7470</u>	.6326	.6068	.5463	.5800	.6612	.5838
Avg. Ranking		1.2	<u>2.0</u>	4.1	5.1	7.3	6.1	3.8	6.4

Table 2: Comparison of different model-agnostic feature attribution methods based on deletion and insertion scores. The results are averaged over 100 target instances \times 100 random trials. DFAX_G and DFAX_S employ GKDE and SiNNE as the kernel density estimator, respectively. ↓ indicates that a lower value for the evaluation metric is better, while ↑ indicates the higher the better. The **best** results are shown in boldface and the second-best results are underlined.

each snippet signifying its sentiment using a feature attribution method. DLIME is selected for comparison with DFAX in this task as it is a leading baseline method in our preceding quantitative experiments. It serves as a representative for the baseline feature attribution methods.

As presented in Table 3, DFAX identifies words such as “compelling”, “fascinating”, and “moving” as most indicative of positive sentiment, while highlighting “bad”, “dull”, and “too” for negative sentiment. In contrast, DLIME fails to identify reasonable words. For instance, it selects “real” over “compelling” for a positive review and “humor” over “bad” for a negative one. These examples suggest the superior quality of the attributions generated by DFAX.

HER2st. Spatial transcriptomics (ST) is a pivotal technology for scientists to profile gene expression and spatial information in tissue samples (Marx 2021). A key application of ST is in oncology, where the over-expression of the human epidermal growth factor receptor 2 (HER2) gene defines the major subtypes of breast cancer. Here we use the HER2st dataset containing HER2-positive breast tumor data collected from an ST platform. This dataset comprises 527 cells, each characterized by a pair of spatial coordinates and the expression values of the 314 most highly variable genes. While previously utilized in clustering (Zhang et al. 2025), this dataset is also well-suited for feature attribution, an application that allows scientists to identify which genes are

most critical in determining a cell’s tissue type.

We train a classifier and obtain initial tissue predictions for each cell, which are visualized in Figure 2(a), plotted using the spatial coordinates provided in the dataset. Then a feature attribution method is employed to identify the 157 (i.e., half the total) most salient genes for each cell. The remaining non-salient genes are then masked with random expression values, and the classifier’s predictions are re-evaluated on these masked cells.¹ The underlying assumption is that an effective attribution method should preserve the most critical genes, thus maintaining high prediction accuracy even after masking (the initial predictions serve as the ground truth labels for this accuracy calculation).

The prediction results for DFAX and DLIME are illustrated in Figures 2(b) and 2(c), respectively. Cells for which the re-evaluated predictions do not match the initial predictions are marked with a cross. As depicted, DFAX achieves a high accuracy of 95.64%, significantly outper-

¹It is important to distinguish this evaluation protocol from the explanatory model building process. Definition 1 requires that the explanatory model be derived from the unmodified dataset \mathbf{X} , a criterion to which our method adheres. The data modification described here is an evaluation protocol used solely to assess any feature attribution methods, after the explanatory models are built. The modification is applied to the testing instances rather than to \mathbf{X} . This evaluation protocol is similar to that used in the deletion and insertion scores reported in the quantitative experiments.

Positive example 1	Positive example 2	Positive example 3
the film’s <i>real</i> 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 <i>documentary</i> that should shame americans. moving film that respects its <i>audience</i> and its source material
Negative example 1	Negative example 2	Negative example 3
contains the <i>humor</i> , characterization, poignancy, and intelligence of a bad sitcom	a dark, dull <i>thriller</i> 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.

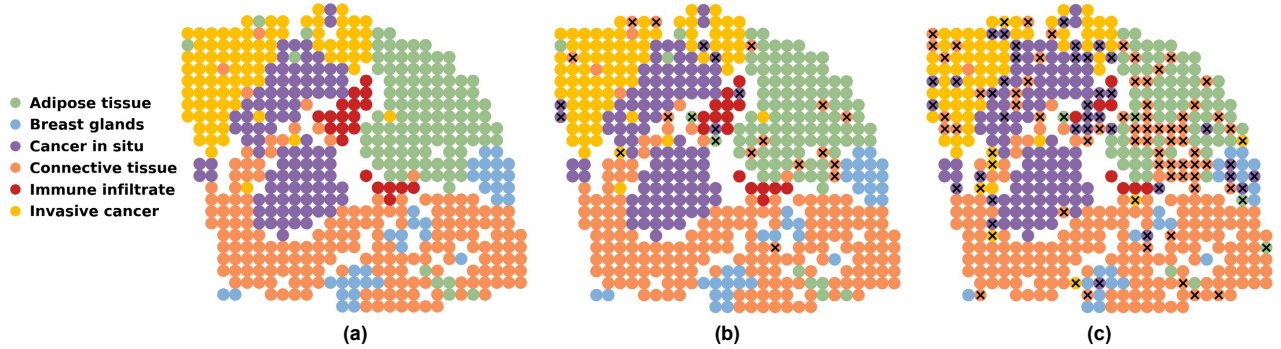


Figure 2: Visualization of tissue predictions on the HER2st dataset. (a) Initial predictions for the original cells with all 314 genes. (b-c) Predictions based on the 157 salient genes only, identified by (b) DFAX and (c) DLIME.

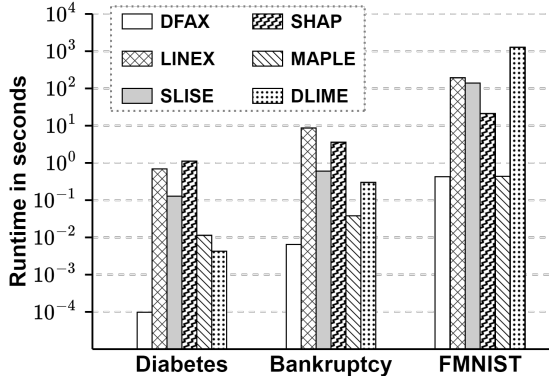


Figure 3: Runtime of the methods in seconds (log-scale).

forming DLIME’s 79.51%. Notably, DFAX results in only 6 misclassifications of cancer cells, compared to 44 for DLIME. These findings strongly demonstrate the effectiveness of DFAX and highlight its potential for real-world scientific applications.

MNIST and FMNIST. Additional qualitative evaluations on these two image datasets can be found in Appendix B.

Runtime Comparison

To demonstrate the efficiency of DFAX, we provide in Figure 3 an empirical runtime comparison on three representative datasets of varying sizes, feature counts, and numbers of classes. It reports the total time required for each method

to generate feature attributions for a single target instance. Figure 3 shows that DFAX achieves the shortest runtime on all three datasets. Notably, DFAX is often faster than other baselines by orders of magnitude. This runtime comparison underscores the efficiency of DFAX, affirming its practicality for generating feature attributions in large-scale real-world applications.

Conclusions

The proposed DFAX is the first fast and effective feature attribution explainer to operate directly from a distributional perspective by means of probability/density estimation. This is made possible via a formal definition for the problem of feature attribution, which reveals the importance of generating explanations supported by distribution \mathcal{P} . Our extensive experiments have demonstrated both quantitatively and qualitatively the superiority of DFAX over other state-of-the-art methods, and the efficiency of DFAX was also confirmed through runtime comparisons.

We contend that many of the limitations found in existing model-agnostic feature attribution methods stem from one underlying issue: they have been designed and evaluated without fully understanding the problem, a direct consequence due to the absence of a formal problem definition. The proposed formal definition addresses this gap, establishing a guideline for the evaluation and design of any explanation methods for feature attribution.

In the near future, we plan to investigate the axiomatic properties of DFAX, and to extend DFAX to the task of feature-group attribution.

Acknowledgments

Kai Ming Ting is supported by the National Natural Science Foundation of China (Grant No. 92470116).

References

- Andersson, A.; Larsson, L.; Stenbeck, L.; Salmén, F.; Ehinger, A.; Wu, S.; Al-Eryani, G.; Roden, D.; Swarbrick, A.; Borg, Å.; et al. 2020. Spatial deconvolution of HER2-positive breast tumors reveals novel intercellular relationships. *bioRxiv*, 2020–07.
- Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82–115.
- Asuncion, A.; Newman, D.; et al. 2007. UCI machine learning repository.
- Bandaragoda, T. R.; Ting, K. M.; Albrecht, D.; Liu, F. T.; and Wells, J. R. 2014. Efficient anomaly detection by isolation using nearest neighbour ensemble. In *2014 IEEE International conference on data mining workshop*, 698–705. IEEE.
- Björklund, A.; Henelius, A.; Oikarinen, E.; Kallonen, K.; and Puolamäki, K. 2023. Explaining any black box model using real data. *Frontiers in Computer Science*, 5: 1143904.
- Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.
- Chen, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.
- Dhurandhar, A.; Natesan Ramamurthy, K.; Ahuja, K.; and Arya, V. 2023. Locally invariant explanations: Towards stable and unidirectional explanations through local invariant learning. *Advances in Neural Information Processing Systems*, 36: 19410–19445.
- Fisher, A.; Rudin, C.; and Dominici, F. 2019. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.
- Freiesleben, T.; and König, G. 2023. Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In *World conference on explainable artificial intelligence*, 48–65. Springer.
- Freund, Y.; and Schapire, R. E. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, 23–37. Springer.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Ivanovs, M.; Kadikis, R.; and Ozols, K. 2021. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150: 228–234.
- Klein, L.; Lüth, C.; Schlegel, U.; Bungert, T.; El-Assady, M.; and Jäger, P. 2024. Navigating the Maze of Explainable AI: A Systematic Approach to Evaluating Methods and Metrics. *Advances in Neural Information Processing Systems*, 37: 67106–67146.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.
- Li, X.; Pan, D.; Li, C.; Qiang, Y.; and Zhu, D. 2023. Negative flux aggregation to estimate feature attributions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 446–454.
- Lipovetsky, S.; and Conklin, M. 2001. Analysis of regression in game theory approach. *Applied stochastic models in business and industry*, 17(4): 319–330.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Marx, V. 2021. Method of the Year: spatially resolved transcriptomics. *Nature methods*, 18(1): 9–14.
- Minh, D.; Wang, H. X.; Li, Y. F.; and Nguyen, T. N. 2022. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 1–66.
- Muandet, K.; Fukumizu, K.; Sriperumbudur, B.; Schölkopf, B.; et al. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2): 1–141.
- Pang, B.; and Lee, L. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, 115–124.
- Petsiuk, V.; Das, A.; and Saenko, K. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.
- Plumb, G.; Molitor, D.; and Talwalkar, A. S. 2018. Model agnostic supervised local explanations. *Advances in Neural Information Processing Systems*, 31.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.
- Rosenblatt, M. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3): 832–837.
- Samariya, D.; Aryal, S.; Ting, K. M.; and Ma, J. 2020. A new effective and efficient measure for outlying aspect mining. In *Web Information Systems Engineering—WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, 2020, Proceedings, Part II 21*, 463–474. Springer.
- Shapley, L. S. 1953. A Value for n-Person Games. In Kuhn, H. W.; and Tucker, A. W., eds., *Contributions to the Theory of Games II*, 307–317. Princeton: Princeton University Press.

Strumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41: 647–665.

Ting, K. M.; Washio, T.; Wells, J. R.; and Zhang, H. 2021. Isolation kernel density estimation. In *2021 IEEE International Conference on Data Mining (ICDM)*, 619–628. IEEE.

Ting, K. M.; Zhu, Y.; and Zhou, Z.-H. 2018. Isolation kernel and its effect on SVM. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2329–2337.

Visani, G.; Bagli, E.; and Chesani, F. 2020. OptiLIME: Optimized LIME explanations for diagnostic computer algorithms. *arXiv preprint arXiv:2006.05714*.

Walker, C.; Jha, S.; Chen, K.; and Ewetz, R. 2024. Integrated decision gradients: Compute your attributions where the model makes its decision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5289–5297.

Wells, J. R.; and Ting, K. M. 2019. A new simple and efficient density estimator that enables fast systematic search. *Pattern Recognition Letters*, 122: 92–98.

Williams, C.; and Seeger, M. 2000. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 13.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Zafar, M. R.; and Khan, N. 2021. Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction*, 3(3): 525–541.

Zhang, H.; Zhang, Y.; Ting, K. M.; Zhang, J.; and Zhao, Q. 2025. Kernel-bounded clustering for spatial transcriptomics enables scalable discovery of complex spatial domains. *Genome Research*, 35(2): 355–367.

Zhu, Z.; Chen, H.; Wang, X.; Zhang, J.; Jin, Z.; Xue, J.; and Shen, J. 2024. Iterative search attribution for deep neural networks. In *Forty-first International Conference on Machine Learning*.