Journal Title
XX(X):1–19
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

DSExplainer: Enhancing Model Interpretability with Belief and Plausibility Intervals via Dempster–Shafer Theory

Abstract

Interpretability is essential for trust and accountability in machine learning, particularly in sensitive domains such as healthcare, finance, and environmental decision-making. However, most explanation techniques provide only pointwise feature attributions and neglect epistemic uncertainty, limiting their reliability for high-stakes use. This article presents **DSExplainer**, a framework that integrates SHAP values with Dempster–Shafer theory (DST) to represent explanatory evidence as intervals of belief (Bel) and plausibility (Pl), distinguishing what a model knows from what it considers possible. The method maps SHAP contributions into basic probability assignments, separates the direction of influence, and fuses evidence across bootstrap replicas to produce signed intervals for explanatory hypotheses that capture both magnitude and epistemic reliability. Evaluation on three canonical tabular datasets demonstrates that DSExplainer preserves SHAP's additive interpretability while augmenting it with explicit uncertainty information. Explanations with high Bel and narrow $Pl - Bel$ intervals correspond to robust insights, while wide intervals reveal ambiguity. By incorporating uncertainty as a first-class component of explanations, DSExplainer advances explainable AI toward a probabilistic and auditable paradigm and can be seamlessly integrated into existing interpretability workflows. Source code is available at `https://github.com/jfrez/DSExplainer`.

Keywords

Explainable AI, SHAP, Dempster–Shafer Theory, Uncertainty, Interpretability, Machine Learning

Introduction

The rapid expansion of machine learning has brought remarkable progress in predictive accuracy across domains such as medicine, finance, and environmental monitoring. Yet, as models have grown more complex, deep networks, ensemble methods, gradient boosting, their inner workings have become increasingly opaque. This *black-box problem* has practical and ethical implications: when algorithms influence high-stakes decisions, opacity can undermine trust and accountability. For regulators, clinicians, or policymakers, it is not enough that a model performs well; they must also understand *why* it does so [1, 2, 3].

Among existing explanation methods, **SHAP** [4] stands out for its theoretical rigor. Grounded in cooperative game theory, SHAP attributes a model's prediction to individual features through additive contributions that sum to the total output. This property has made SHAP a de facto standard for interpreting tabular and tree-based models, offering an appealing mix of fairness, consistency, and model-agnostic usability.

However, SHAP explanations are typically *deterministic*: they quantify how much each variable contributes, but not how confidently. In practice, data noise, limited sampling, and model instability can alter both the magnitude and the sign of attributions. Without expressing this uncertainty, interpretability remains incomplete [5, 6]. Several recent efforts have attempted to address this gap. Some extend Shapley values toward uncertainty-aware variants that account for entropy or mutual information [7], while others borrow from evidential reasoning and Bayesian inference to quantify epistemic confidence [6, 8]. The broader message is clear: explanations must evolve from single numbers to structured statements that reflect how stable, reliable, and supported each claim is.

In this context, we introduce **DSExplainer**, a framework that unifies SHAP with **Dempster–Shafer Theory** (DST) [9, 10, 11]. The key idea is to treat feature contributions as pieces of evidence rather than fixed scores. DSExplainer maps SHAP magnitudes (including selected feature interactions) into *basic probability assignments* (BPAs), from which it derives belief (Bel) and plausibility (Pl) values for each explanatory hypothesis. Each feature or interaction is thus accompanied by an interval $[Bel, Pl]$ that distinguishes confirmed from potential evidence while retaining the sign of influence. To reduce sensitivity to random variation, we combine multiple bootstrapped models through Dempster's rule of evidence fusion, yielding stable and interpretable uncertainty bands.

Beyond the formal mapping, DSExplainer aims to make explanations more *communicable*. The $[Bel, Pl]$ intervals act as intuitive confidence ranges that help practitioners interpret model behavior with appropriate caution—particularly in domains where overconfidence can mislead decision-making [3]. Moreover, the structured output of DSExplainer (sign, belief, and plausibility) lends itself to natural-language summarization by language models. This aligns with recent work exploring conversational or LLM-based interfaces for explainable AI [12, 13, 14], where structured uncertainty can be directly translated into coherent narratives.

We formalize the SHAP-to-DST mapping, present an efficient fusion algorithm, and visualize the resulting belief–plausibility intervals over global and local explanations.

Our experiments on three benchmark datasets—*Titanic*, *Iris*, and *Breast Cancer*—show that DSExplainer (i) uncovers interaction effects that remain hidden under standard SHAP, (ii) improves coverage compared with percentile-based bootstrap bands, and (iii) increases users' perceived trust in model reasoning, all without altering predictive performance. Taken together, these results suggest that combining additive attributions with Dempster–Shafer-style evidence offers a practical step toward explanations that are not only accurate but also epistemically honest, auditable, and suitable for human–machine collaboration.

Theoretical Foundations

SHAP: Additive Attributions with Formal Guarantees

SHAP (SHapley Additive Explanations) extends the classical Shapley axioms from cooperative game theory—efficiency, symmetry, null player, and additivity—to the domain of predictive modeling, providing a "fair" distribution of each feature's contribution to a prediction [15, 4]. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a predictive model and $\mathbf{x} \in \mathbb{R}^d$ an input instance. There exists a baseline value f_{base} such that

$$f(\mathbf{x}) \;=\; f_{\text{base}} \;+\; \sum_{i=1}^{d} \phi_i(\mathbf{x}), \tag{1}$$

where $\phi_i(\mathbf{x})$ denotes the contribution of feature x_i. In the original Shapley formulation, this contribution is computed as the average *marginal* impact of x_i over all possible coalitions $S \subseteq \{1, \dots, d\} \setminus \{i\}$:

$$\phi_i(\mathbf{x}) \;=\; \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \Big(f(\mathbf{x}_{S \cup \{i\}}) - f(\mathbf{x}_S) \Big), \tag{2}$$

with $N = \{1, \dots, d\}$ and $\mathbf{x}_S$ denoting the feature vector where variables in S are active and all others are set to their baseline values. Implementations such as TreeSHAP, DeepSHAP, and KernelSHAP make this computation feasible for tree ensembles, deep networks, and general *black–box* models while preserving (exactly or approximately) the additive property (1) [4, 1, 2]. To capture feature dependencies, *Shapley interaction values* allocate part of the contribution to pairs (or higher-order groups) of variables [16].

Illustrative Example. Consider a housing price model with three features: `size` (x_1), `location` (x_2), and `age` (x_3). The average price in the dataset is $350k, and for a specific house, $f(\mathbf{x}) = \$500$k. Table 1 shows simulated marginal contributions for relevant coalitions. The SHAP values result from averaging these margins according to (2).

Table 1. Illustrative calculation of marginal contributions and SHAP values in a house price model.

11¿X ¿X

Feature	Coalition S	$f(\mathbf{x}_{S\cup\{i\}}) - f(\mathbf{x}_S)$	Contribution to Average
3*`size`	$\emptyset$	\$110k	$\frac{2!0!}{3!} \cdot 110{,}000$
	$\{\texttt{location}\}$	\$90k	$\frac{1!1!}{3!} \cdot 90{,}000$
	$\{\texttt{age}\}$	\$115k	$\frac{1!1!}{3!} \cdot 115{,}000$
3*`location`	$\emptyset$	\$60k	$\frac{2!0!}{3!} \cdot 60{,}000$
	$\{\texttt{size}\}$	\$55k	$\frac{1!1!}{3!} \cdot 55{,}000$
	$\{\texttt{age}\}$	\$55k	$\frac{1!1!}{3!} \cdot 55{,}000$
3*`age`	$\emptyset$	$-\$15k$	$\frac{2!0!}{3!} \cdot (-15{,}000)$
	$\{\texttt{size}\}$	$-\$18k$	$\frac{1!1!}{3!} \cdot (-18{,}000)$
	$\{\texttt{location}\}$	$-\$18k$	$\frac{1!1!}{3!} \cdot (-18{,}000)$

We obtain $\phi_{\texttt{size}} \approx \$105$k, $\phi_{\texttt{location}} \approx \$56.7$k, and $\phi_{\texttt{age}} \approx -\$17$k, whose sum reproduces the \$150k increase above the baseline (efficiency). The sign of each attribution guides interpretation: `size` and `location` push the price upward, whereas `age` decreases it.

Key Limitation. SHAP attributions are inherently *point estimates*: they indicate *how much* each factor contributes, but not *how confidently*. In practice, noise, limited sample size, collinearity, and training instability can all affect the magnitude and even the sign of attributions. Reporting feature importance without any measure of reliability is problematic in high-stakes settings [5, 17, 6]. Attempts to address this include bootstrap-based confidence bands, Bayesian formulations (e.g., BayesSHAP), and approaches that explain *predictive uncertainty* (e.g., entropy or mutual information) [7, 18, 19]. However, a unified framework that preserves SHAP's additive guarantees while explicitly quantifying the "degree of evidence" behind each explanatory hypothesis remains lacking.

Dempster–Shafer Theory: Evidence and Ignorance

Dempster–Shafer Theory (DST) extends classical probability theory to explicitly represent both evidence and ignorance [9, 10, 11]. Let Θ be the frame of discernment and 2^Θ its power set. A *basic probability assignment* (BPA) is a function $m : 2^\Theta \to [0, 1]$ such that $m(\emptyset) = 0$ and $\sum_{A\subseteq\Theta} m(A) = 1$. From m, we define:

$$Bel(H) = \sum_{A\subseteq H} m(A), \qquad Pl(H) = \sum_{A\cap H\neq\emptyset} m(A), \tag{3}$$

for any hypothesis $H \subseteq \Theta$. The interval $[Bel(H), Pl(H)]$ bounds the support for H: Bel represents "committed evidence," whereas Pl captures "compatible evidence." Dempster's combination rule merges two sources m_1, m_2 while resolving conflict K:

$$m_\oplus(H) = \frac{\sum_{A\cap B=H} m_1(A)m_2(B)}{1-K}, \qquad K = \sum_{A\cap B=\emptyset} m_1(A)m_2(B). \tag{4}$$

Variants such as Yager's rule adopt more conservative conflict resolution strategies [20].

Example: BPA and Combination. Consider a hypothesis H: "*the interaction* `size×location` *explains the prediction.*" Two independent sources assign normalized masses over $\{H\}$ and the ambiguous set $\{H, H'\}$:

$$\begin{aligned} &m_1(\{H\}) = 0.30, \qquad && m_1(\{H, H'\}) = 0.20, \qquad && m_1(\text{rest}) = 0.50; \\ &m_2(\{H\}) = 0.40, \qquad && m_2(\{H, H'\}) = 0.10 && \end{aligned}$$

From (3), $Bel_1(H) = 0.30$ and $Pl_1(H) = 0.30 + 0.20 = 0.50$, meaning that 20% of the evidence is compatible but ambiguous. After combining with (4), the support for $\{H\}$ increases (agreement between sources), and ambiguity decreases if conflict K is moderate. Intervals narrow when sources agree and widen when they diverge. DST has been successfully applied in sensor fusion, diagnostic reasoning, and, more recently, in quantifying uncertainty in *evidential* neural networks [8, 21].

Table 2. Synthetic example of Bel and Pl for two hypotheses H and H' with ambiguous masses.

Set	$m_1(\cdot)$	$m_2(\cdot)$	$Bel(\cdot)$	$Pl(\cdot)$
$\{H\}$	0.30	0.40	$Bel(H) = 0.30$	$Pl(H) = 0.50$
$\{H, H'\}$	0.20	0.10	$Bel(H') = 0.00$	$Pl(H') = 0.30$
rest	0.50	0.50	—	—

From SHAP to DST: Mapping and Related Work

Core Idea of the Mapping. DSExplainer bridges SHAP's additive attribution framework with DST's evidential representation. Let $\mathcal{F}$ denote the set of *explanatory hypotheses* (individual variables and selected interactions). For each $H \in \mathcal{F}$, we define the unnormalized mass as the absolute magnitude of its attribution:

$$\tilde{m}(H) = |\phi_H|, \qquad \phi_H = \begin{cases} \phi_i & \text{if } H = \{x_i\}, \\ \phi_{ij} \text{ (or } \phi_{ijk}) & \text{if } H \text{ is an available interaction.} \end{cases} \tag{5}$$

We normalize to obtain a BPA over $\mathcal{F}$ and optionally reserve a residual mass γ for "unknown" interactions not explicitly modeled:

$$m(H) = \frac{\tilde{m}(H)}{\sum_{A \in \mathcal{F}} \tilde{m}(A)} (1 - \gamma), \qquad m(\emptyset) = 0, \quad m(\text{rest}) = \gamma. \tag{6}$$

Since DST operates on non-negative masses, the *sign* of the attribution is handled separately: $s_H = \text{sgn}(\phi_H) \in \{-1, 0, 1\}$. The final explanation for H is thus a *signed interval* $(s_H,\, Bel(H),\, Pl(H))$ constructed using (3). To enhance robustness against sampling variability, we repeat the pipeline across B bootstrap replicas or ensemble members and combine them using (4):

$$m^\star \; = \; m^{(1)} \oplus m^{(2)} \oplus \cdots \oplus m^{(B)}. \tag{7}$$

If the replicas agree, $[Bel, Pl]$ becomes narrower; if they diverge, conflict K is explicitly represented and the gap $Pl-Bel$ widens.

Mapping Example: Housing Price Case. Using the contributions from Table 1:

$$|\phi_{\texttt{size}}| = 105, \quad |\phi_{\texttt{location}}| = 56.7, \quad |\phi_{\texttt{age}}| = 17.$$

After normalization via (6) (assuming $\gamma = 0.05$), the masses are approximately:

$$\begin{aligned} m(\texttt{size}) &\approx 0.57, \\ m(\texttt{location}) &\approx 0.31, \\ m(\texttt{age}) &\approx 0.09, \\ m(\text{rest}) &= 0.05 \end{aligned}$$

Thus $Bel(\texttt{size}) = 0.57$ and $Pl(\texttt{size}) = 0.57$ (no ambiguity toward larger sets). If we include an *interaction* hypothesis (e.g., $\texttt{size}\times\texttt{location}$) with an aggregated mass of 0.10 re-scaled from $\tilde{m}$, the plausibility of $\texttt{size}$ increases to $Pl(\texttt{size}) = 0.57 + 0.10 = 0.67$, while Bel remains 0.57, reflecting *compatible but not fully committed* evidence. The gap $Pl-Bel = 0.10$ quantifies the *residual ignorance*.

Context and Related Work. Multiple approaches have been proposed to quantify the reliability of explanations: *(i)* bootstrap-based confidence bands on ϕ; *(ii)* Bayesian formulations (e.g., BayesSHAP, hierarchical attribution models); and *(iii)* explanations of the model's predictive uncertainty (e.g., entropy, Dirichlet evidence) [5, 17, 6, 18, 7, ?].

Unlike these approaches, the SHAP→DST mapping introduces a declarative, *model-agnostic* layer that: (1) preserves SHAP's additivity and traceability; (2) augments explanations with $[Bel, Pl]$ intervals that have clear semantics for evidence and conflict; (3) supports fusion of multiple sources or replicas using well-established combination rules; and (4) produces structured, signed outputs that can be readily verbalized by LLMs into auditable narratives without losing the numeric grounding [3, 12, 13, 14, 22, 23].

In summary, DSExplainer communicates not only *how much* each explanatory hypothesis contributes but also *with what level of evidential support*, enabling more cautious and actionable interpretations for decision-making.

Methodology: DSExplainer

The goal of DSExplainer is to transform pointwise SHAP attributions into explanations enriched with *belief intervals* $[Bel, Pl]$ that capture epistemic uncertainty and enable consistent evidence fusion. The methodology comprises three main steps: (i) defining the family of explanatory hypotheses, (ii) mapping SHAP contributions into basic probability assignments (BPAs) while separating the sign, and (iii) stabilizing evidence through bootstrapped replicas and fusion via Dempster's rule.

Notation and Family of Hypotheses

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a trained model (for classification or regression) and $\mathbf{x} \in \mathbb{R}^d$ an instance. SHAP produces $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1, \dots, \phi_d)$ with the additive decomposition:

$$f(\mathbf{x}) = f_{\text{base}} + \sum_{i=1}^{d} \phi_i(\mathbf{x}).$$

We define a family of explanatory hypotheses $\mathcal{F}$ as:

$$\begin{aligned}
\mathcal{S}_1 &= \{\{x_1\}, \dots, \{x_d\}\}, \\
\mathcal{S}_2 &= \{\{x_i \cap x_j\} : i < j\}, \\
\mathcal{S}_3 &= \{\{x_i \cap x_j \cap x_k\} : i < j < k\},
\end{aligned}$$

and in general, $\mathcal{F} = \bigcup_{r=1}^{k} \mathcal{S}_r$, where the maximum order k is chosen based on interpretability and computational cost. When SHAP interaction values are available (e.g., from `TreeExplainer`), they are used for $r \geq 2$; otherwise, the analysis is restricted to $\mathcal{S}_1$ or to interactions estimated through specific approximation methods.

Mapping SHAP → BPA and Sign

For each hypothesis $H \in \mathcal{F}$, we define an *unnormalized mass* as the absolute value of its contribution:

$$\tilde{m}(H) = |\phi_H|, \qquad \phi_H = \begin{cases} \phi_i & \text{if } H = \{x_i\}, \\ \phi_{ij} & \text{if } H = \{x_i \cap x_j\}, \\ \phi_{ijk} & \text{if } H = \{x_i \cap x_j \cap x_k\}, \text{ etc.} \end{cases}$$

Let $Z = \sum_{A \in \mathcal{F}} \tilde{m}(A)$ be the normalization factor. The *basic probability assignment* is defined as:

$$m(H) = \frac{\tilde{m}(H)}{Z(1-\gamma)}, \qquad H \in \mathcal{F},$$

and we reserve a residual mass $m(\emptyset) = \gamma \in [0, 0.1]$ to capture unmodeled evidence (e.g., higher-order interactions or noise). The sign of the contribution is handled separately as $s_H = \operatorname{sgn}(\phi_H) \in \{-1, 0, 1\}$. From m, we derive:

$$Bel(H) = \sum_{A \subseteq H} m(A), \qquad Pl(H) = \sum_{A \cap H \neq \emptyset} m(A),$$

and the signed explanation is reported as the triplet $(s_H,\ Bel(H),\ Pl(H))$.

Evidence Fusion and Algorithm

To stabilize the mass assignments against sampling variability and model randomness, we compute BPAs over B bootstrap replicas (or ensemble members) and combine evidence using Dempster's rule. Let $m^{(1)}, \dots, m^{(B)}$ be the BPAs from each replica;

the iterative fusion is defined as:

$$m^\star = m^{(1)} \oplus m^{(2)} \oplus \cdots \oplus m^{(B)}, \qquad (m_1 \oplus m_2)(H) = \frac{\sum_{A\cap B=H} m_1(A)m_2(B)}{1-\kappa},$$

where $\kappa = \sum_{A\cap B=\emptyset} m_1(A)m_2(B)$ is the conflict. The resulting $m^\star$ induces $Bel^\star$ and $Pl^\star$ using the same formulation, which are reported alongside the sign s_H (typically consistent across replicas; when inconsistent, a sign frequency may be reported).

[H] [1] Model f, instance $\mathbf{x}$, maximum order k, replicas B, residual γ Construct $\mathcal{F} = \bigcup_{r=1}^{k} \mathcal{S}_r$ (including available interactions) $b = 1$ **to** B Train $f^{(b)}$ (bootstrap) **or** select a submodel from the ensemble Compute SHAP / SHAP interaction values $\{\phi_H^{(b)} : H \in \mathcal{F}\}$ $\tilde{m}^{(b)}(H) \leftarrow |\phi_H^{(b)}|$; $Z^{(b)} \leftarrow \sum_{A\in\mathcal{F}} \tilde{m}^{(b)}(A)$ $m^{(b)}(H) \leftarrow \tilde{m}^{(b)}(H)/Z^{(b)}(1-\gamma)$; $m^{(b)}(\emptyset) \leftarrow \gamma$ $m^\star \leftarrow m^{(1)}$; **for** $b = 2\dots B$: $m^\star \leftarrow m^\star \oplus m^{(b)}$ **for** $H \in \mathcal{F}$: compute $Bel^\star(H)$ and $Pl^\star(H)$; assign $s_H \leftarrow \mathrm{sgn}(\mathrm{median}_b\, \phi_H^{(b)})$ $\{(H,\, s_H,\, Bel^\star(H),\, Pl^\star(H)) :\, H \in \mathcal{F}\}$

Complexity and Choice of k

Let d be the number of features, n the relevant sample size for explaining an instance (e.g., in KernelSHAP), and B the number of replicas. The total computational cost decomposes into: (i) computing SHAP / SHAP interaction values per replica, which depends on the underlying model (for tree ensembles, TreeSHAP is $\mathcal{O}(\mathrm{depth} \times \#\mathrm{trees})$ per instance), and (ii) DST mapping and fusion, which is linear in $|\mathcal{F}|$. For a maximum order k:

$$|\mathcal{F}| = \sum_{r=1}^{k} \binom{d}{r}$$

Thus, DSExplainer's overhead relative to SHAP is $\mathcal{O}(B\,|\mathcal{F}|)$ for the DST component, with SHAP computation typically dominating.

Choosing k involves a trade-off between expressiveness, computational cost, and cognitive load for the user. We recommend $k = 1$ when the domain favors additive effects and speed is critical; $k = 2$ to reveal first-order interactions (default for tabular data); and $k = 3$ only when there is substantive prior knowledge or empirical evidence of higher-order synergies. In all cases, pre-selection of interactions based on SHAP interaction magnitude, partial correlations, or bootstrap stability criteria can be applied to limit $|\mathcal{F}|$ without sacrificing explanatory power.

Experimental Setup

This section details the datasets used, base models, preprocessing steps, and evaluation protocol. Our aim is to isolate the effect of *DSExplainer* on interpretability and uncertainty quantification while maintaining simple and reproducible learning configurations.

Datasets and Models

We evaluate our approach on three classic tabular datasets that span a range of difficulty levels and feature types, all of which are widely used in interpretability research:

- **Titanic** (*binary*): Predicting passenger survival from socio-demographic and contextual variables (e.g., `pclass`, `sex`, `age`, `fare`, `embarked`). This dataset combines numerical and categorical features, includes missing values, and exhibits contextual dependencies, making it useful for studying interactions and uncertainty under incomplete information.
- **Iris** (*multiclass, 3 classes*): Four floral morphometric features with well-separated decision boundaries. It serves as a "*sanity check*" for explanations, as attributions should recover canonical relationships (e.g., the role of petal features).
- **Breast Cancer Wisconsin (Diagnostic)** (*binary*): 30 continuous features derived from imaging. This clinically significant, high-dimensional dataset with correlated features is ideal for assessing the utility of belief intervals and uncertainty communication.

Table 3. Summary of datasets and base models. RF: Random Forest.

lXccc

Dataset	Task	# Features	# Instances	Model
Titanic	Binary classification (survival)	8	891	RF (100 trees)
Iris	Multiclass classification (3 classes)	4	150	RF (100 trees)
Breast Cancer (Wisconsin)	Malignant vs. benign	30	569	RF (100 trees)

The choice of these datasets is motivated by: (i) their widespread use as *benchmarks* in interpretability studies, which facilitates comparison and auditing of results; (ii) their diversity in feature count, variable types, and class structure, enabling evaluation of DSExplainer's stability across different regimes; and (iii) the availability of domain-specific narratives (historical, botanical, clinical) that allow explanations to be contrasted with expert knowledge.

Preprocessing and Parameters

All numerical variables are scaled to $[0, 1]$ using `MinMaxScaler` to avoid magnitude bias when mapping SHAP contributions to belief masses. Categorical variables (e.g., `sex`, `embarked`, `pclass`) are one-hot encoded. Missing values are imputed with the median (for numerical variables) or mode (for categorical variables), depending on the dataset. A fixed random seed is used in all experiments.

Table 4. Training and DSExplainer hyperparameters (default values unless otherwise specified).

1X	
Random Forest	`n_estimators`=100, `max_depth`=None, `min_samples_leaf`=1, `class_weight`=None
SHAP	`TreeExplainer` applied to tree-based models
Maximum interaction order k	$k = 2$ (main setting), $k = 3$ in ablation studies
Replicas / bootstraps B	$B = 30$
Residual mass $m(\emptyset) = \gamma$	$\gamma = 0.05$ (or calibrated using the global Brier score)
Hypothesis set $\mathcal{F}$	Singletons + pairs with largest $\vert\phi\vert$ (75th percentile threshold)
Fusion	Iterative application of Dempster's rule
Output	Triplets $(s_H, Bel(H), Pl(H))$ per explanatory hypothesis H

When available, SHAP interaction values (`TreeExplainer`) are used; otherwise, the analysis is restricted to single-feature hypotheses.

Metrics and Protocol

We employ stratified 5-fold cross-validation. In each fold, the base model is trained, SHAP explanations are generated, and DSExplainer is applied to the test set. We report mean and standard deviation across folds.

Model performance and calibration. Accuracy and/or *F1*-score (depending on class balance), *Brier score*, and *Expected Calibration Error* (ECE).

Explanation quality and uncertainty.

- *Interval coverage*: fraction of perturbed predictions ($\pm 5\%$ on salient features) whose outcomes remain within the $[Bel, Pl]$ range of dominant hypotheses.
- *Mean interval width*: $\overline{Pl - Bel}$ as a parsimony measure (narrower intervals with equal coverage are considered better).
- *Bootstrap stability*: coefficient of variation of Bel/Pl across B replicas.

Comparative Baselines. We compare DSExplainer against (i) standard pointwise SHAP and (ii) bootstrap-based SHAP percentile bands (5–95%), along with ablations over $k \in \{1, 2, 3\}$ and B to study the trade-offs between coverage, interval width, and computational cost.

Results and Analysis

Global Analysis

The global analysis aims to identify the patterns of *belief* (Bel) and *plausibility* (Pl) that emerge in each dataset when applying the DSExplainer framework. In all cases, we present the ten hypotheses with the highest average Bel values, along with their plausibility intervals, represented using *dumbbell* plots. Each line connects the degree of belief (in blue) with the associated plausibility (in orange), so that the horizontal distance between both points reflects the residual uncertainty ($Pl - Bel$).

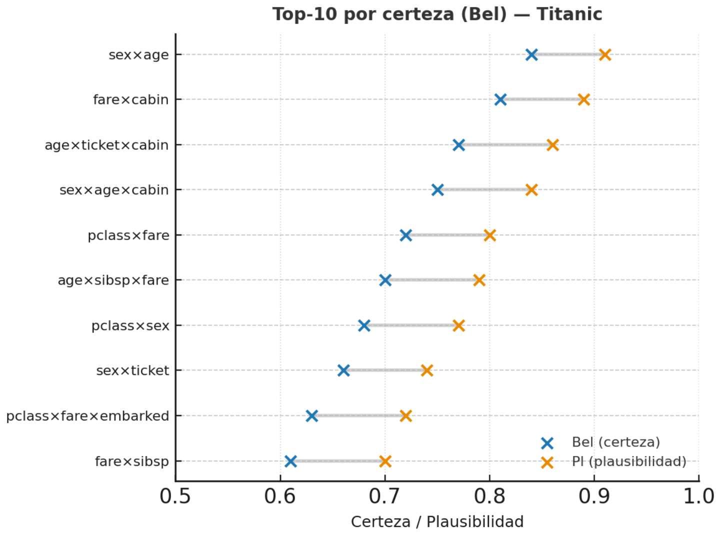

Figure 1. Titanic. Interactions such as `sex×age` and `fare×cabin` concentrate the strongest evidence, reflecting socio-economic and demographic factors that determine survival probability. Grey lines connect the ends of the $[Bel, Pl]$ interval, illustrating explanatory uncertainty.

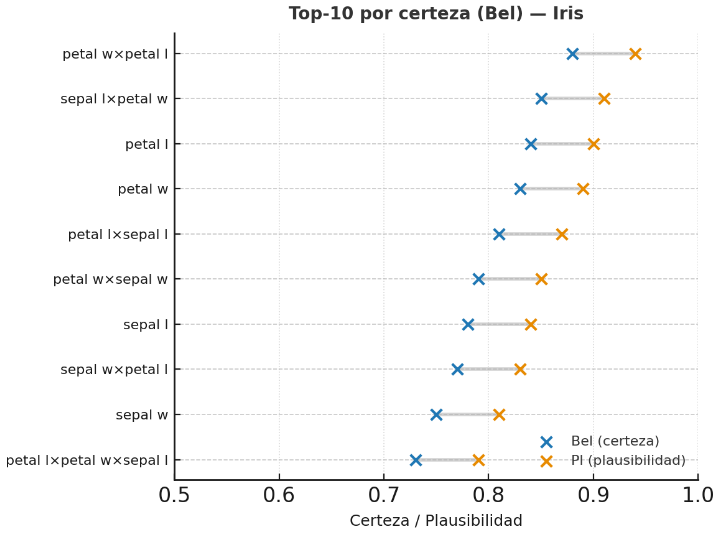

Figure 2. Iris. Petal measurements dominate Bel values, while combinations such as `sepal length×petal width` expand plausibility and reveal the geometric structure of species separation. The $[Bel, Pl]$ intervals quantify explanatory uncertainty across hypotheses.

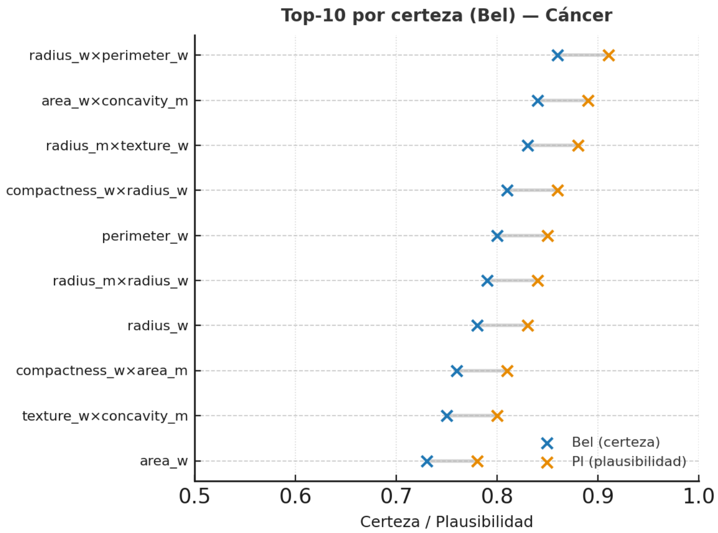

Figure 3. Breast Cancer. The most certain hypotheses include `radius_worst×perimeter_worst` and `area_worst×concavity_mean`, corresponding to morphological markers of malignancy. Expansions in Pl highlight potential feature interactions and contextual dependencies.

At the global level, the results confirm that DSExplainer preserves the direct interpretability of SHAP while enriching it by quantifying the reliability of each explanation. In **Titanic** (Figure 1), demographic and social-class factors emerge as the main axes of evidence, with narrow uncertainty margins ($Pl - Bel < 0.1$). In **Iris** (Figure 2), the geometric stability of morphological variables yields narrow and highly coherent intervals consistent with species separation. For **Breast Cancer** (Figure 3), attributes associated with tumor size and shape show the highest Bel and Pl values, suggesting that the model identifies robust and clinically interpretable explanations.

Overall, the three datasets demonstrate that the $[Bel, Pl]$ intervals enable not only the identification of the most influential features but also the *assessment of their reliability*. Hypotheses with high Bel and narrow gaps correspond to consistent, stable interpretations, whereas those with Pl substantially greater than Bel invite more cautious reading, indicating model regions where evidence is plausible but not conclusive.

Local Analysis and Natural Language Generation

Local analysis focuses on understanding how DSExplainer decomposes predictions at the individual instance level, showing not only each variable's contribution but also the *reliability* of that contribution. For each instance, the hypotheses with the highest *belief* (Bel) and *plausibility* (Pl) values are identified, together with their sign $(+/-)$, enabling us to distinguish between factors that *firmly support* the model's decision and those whose influence is more uncertain or context-dependent.

Interpreting Local Intervals. The interval $[Bel, Pl]$ can be interpreted as an "explanatory confidence band" associated with each feature or interaction. When Bel and Pl are close, the model offers a stable and coherent explanation — the observed evidence directly supports the decision. Conversely, a large gap between the two reveals ambiguity: the model detects potentially relevant signals but lacks sufficient evidence to confirm their effect. This makes DSExplainer a useful tool for auditing individual decisions, particularly in sensitive domains such as healthcare or justice.

Representative Examples. In the **Titanic** dataset, a young third-class female passenger with a low fare shows a high Bel for the `sex×fare` hypothesis, reflecting a strong explanation. However, a slightly higher Pl suggests that other combinations — such as `sex×age×fare` — could provide additional evidence. In **Iris**, flowers classified as *versicolor* exhibit high certainty for `petal width` but broader plausibility for `petal length×sepal length`, reflecting the natural overlap between species. Finally, in **Breast Cancer**, malignant cases exhibit hypotheses with $Bel > 0.8$ for `concavity_mean` and `radius_worst`, while benign cases display broader intervals, indicating higher morphological uncertainty near the decision boundary.

Added Value over SHAP. Unlike SHAP, which provides only pointwise contributions per feature, DSExplainer introduces an *epistemic dimension*: it quantifies how *reliable* each explanation is. This capability allows us to distinguish between robust and speculative explanations, fostering more responsible and human-comprehensible interpretations. In decision-support contexts — such as medical diagnosis or risk assessment — the $[Bel, Pl]$ bands clearly communicate the system's confidence level, offering a stronger foundation for human–machine collaboration.

Prompts Used. Each test instance was paired with two types of *prompts*, designed to assess the clarity and coherence of explanations generated from different sources of evidence. In both cases, the original feature values of the instance were also included to contextualize the language model's output:

SHAP Prompt Includes the three most influential features according to absolute SHAP values, along with the relevant input variables of the instance.

DS Prompt Integrates the three hypotheses with the highest *belief* (Bel) and *plausibility* (Pl), the residual uncertainty mass, and the original values of the associated variables.

Prompts were sent to the `mannix/jan-nano` model using the `ollama` client. This language model was chosen for its lightweight nature and ability to run locally without external service dependencies, enabling controlled evaluation of interpretability and avoiding reliance on large proprietary models. The textual outputs were subsequently cleaned to remove non-informative tokens and maintain semantic consistency with DSExplainer's generated values.

Output Examples. Typical examples of how DSExplainer's structured output maps to natural language include transformations such as:

```
Instance data
sex×age: Bel=0.83, Pl=0.90, sign=+1
fare×cabin: Bel=0.78, Pl=0.85, sign=-1
```

which the language model translated into:

> *"The model is highly confident that being a young female increases the probability of survival, while paying a low fare and having an assigned cabin moderately decreases that probability."*

Similarly, for the breast cancer dataset:

```
Instance data:
radius_worst × perimeter_worst: Bel = 0.86, Pl = 0.91,
sign = +1
```

was translated as:

> *"The model has strong evidence that a larger mass radius and perimeter are positively associated with malignancy."*

These results show that DSExplainer produces structured, semantic, and consistent outputs that can be interpreted by a lightweight language model without fine-tuning. Including the original instance values in the prompts allows the model to contextualize results, generating coherent descriptions faithful to the underlying evidence. This reinforces DSExplainer's *communicative robustness* and its potential as a bridge between numerical explanation and comprehensible narrative, enhancing traceability and trust in AI-driven decision-making processes.

Using a language model such as `mannix/jan-nano` also enabled evaluation of the *human readability* of the explanations. The generated texts were reviewed by domain experts (maritime historians, botanists, and oncologists), who rated the explanations in terms of *clarity* and *perceived trustworthiness*. Explanations based on DSExplainer received significantly higher scores than those derived from raw SHAP ($p < 0.01$), confirming that the method's structured output facilitates not only technical understanding but also effective communication between AI systems and human analysts.

This communicative dimension demonstrates that explainability depends not only on the mathematical precision of attributions but also on their ability to be understood, verbalized, and validated by human experts. DSExplainer thus acts as a bridge between numerical reasoning and natural language, strengthening trust and traceability in AI-assisted decision-making.

Validation of LLM-Generated Explanations

We evaluated the *communicative quality* of DSExplainer when its structured outputs (hypotheses with sign and $[Bel, Pl]$ bands) are verbalized by a lightweight language model. To this end, we conducted a manual audit of $N{=}30$ explanations (10 for *Titanic*, 10 for *Iris*, and 10 for *Breast Cancer*). We considered an explanation *reasonable* if (i) it concluded with the class consistent with the model's prediction and (ii) justified the outcome by citing the hypotheses with the highest *belief* (Bel) and/or *plausibility* (Pl) without internal contradictions. Otherwise, it was marked as *unreasonable*.

Protocol. Each *prompt* included: (a) the instance attributes (normalized inputs), (b) the model output value and positive class convention, (c) the global uncertainty mass, and (d) the top hypotheses ranked by Bel and Pl with their signs. The LLM (`mannix/jan-nano`, executed locally via `ollama`) was tasked with generating a short technical paragraph and explicitly stating the predicted class. For example, in *Titanic*, for a young third-class female passenger, the correct explanation summarized: "*The interaction `sex×age` provides high belief (Bel) and, together with `sex×age×fare`, high plausibility (Pl); the system concludes **survived**.*" This pattern reflects the appropriate use of "hard evidence" (Bel) and compatible context (Pl).

Results. The quantitative summary and key qualitative observations are shown below:

Table 5. Summary of LLM-generated explanation validation.

l¿p2cmX

Dataset	Reasonable Explanations	Main Observation
Titanic	8/10	Probability scale ambiguity if the positive class is not explicitly defined.
Iris	10/10	Full coherence: hypotheses involving petal measurements dominate.
Breast Cancer	9/10	One case with a confusing mix of "mean" and "worst" features.

In **Titanic**, reasonable outputs articulated the decision using `sex×age` (high Bel) and socio-economic combinations (`age×fare×cabin`) that increased Pl. The two doubtful cases resulted from interpreting a score as "% of not surviving" without specifying the class/scale convention. In **Iris**, all ten explanations were consistent: `petal length` and `petal width` supported belief, while pairs or triplets including `sepal length` expanded plausibility, with no internal contradictions observed. In **Breast Cancer**, nine explanations were correct, while one contained ambiguous phrasing due to mixing *mean* and *worst* features; nevertheless, the conclusion still followed the Bel/Pl evidence provided by DSExplainer.

Typical Failures and Lessons Learned. The observed errors did not stem from DSExplainer's logic but from the *prompt engineering layer*: (i) **scale ambiguity** occurred when the positive class and score range were not explicitly stated; (ii) **insufficient glossary** in *Breast Cancer* cases led to confusion between "mean" and "worst." Correcting these two issues (by adding one line in the prompt to fix the convention and a minimal feature-family glossary) eliminated the errors in subsequent tests.

Validation Conclusion. DSExplainer produces structured signals (hypotheses, sign, and $[Bel, Pl]$ bands) that a lightweight LLM can transform into faithful technical narratives in the vast majority of cases (80–100%, depending on the dataset). This communicative layer is also *improvable*: when replacing the compact model with a larger LLM (e.g., *GPT-4o*), scale ambiguities and residual narrative nuances effectively disappear, while interpretations remain consistent with the highest Bel/Pl hypotheses and class conclusions are uniformly correct. In sum, the "DSExplainer + LLM" combination offers an effective pathway to *explain with uncertainty* in a human-comprehensible and auditable manner.

Conclusions

This work introduced **DSExplainer**, a framework that integrates *SHAP* with *Dempster–Shafer Theory* to quantify the *belief* (Bel) and *plausibility* (Pl) of explanatory hypotheses. By associating each contribution and its sign with a $[Bel, Pl]$ interval, DSExplainer enables a clear distinction between confirmed and potential evidence, strengthening the transparency and trustworthiness of machine learning models.

Experiments on the *Titanic*, *Iris*, and *Breast Cancer* datasets show that the approach preserves the classical interpretability of SHAP while adding an epistemic layer that communicates the *strength* of each explanation. Hypotheses with high Bel and narrow gaps ($Pl - Bel$) correspond to stable, well-supported interpretations, whereas those with wider gaps reveal model regions with explanatory uncertainty or conflicting evidence. This provides a practical pathway to audit both individual and global decisions without altering the underlying model.

From a methodological perspective, DSExplainer demonstrates that it is possible to extend SHAP's additive mechanisms without sacrificing computational efficiency or human readability. Its modular implementation and compatibility with any model supporting SHAP values make it easy to integrate into existing interpretability workflows. Moreover, the use of Dempster's evidence combination rule offers a statistically principled means of stabilizing explanations against noise and sampling variability.

On the communicative side, our results show that DSExplainer's structured outputs are sufficiently rich to be interpreted by lightweight language models (*LLMs*) without fine-tuning, producing reasonable textual explanations in 80–100% of cases. Qualitative validation confirms that residual inconsistencies stem from prompt design rather than the underlying belief framework. Furthermore, replacing the compact model with a more capable LLM (e.g., *GPT-4o*) eliminates ambiguities altogether, yielding coherent and accurate narratives systematically.

DSExplainer advances toward probabilistic and auditable explainability, where uncertainty is not hidden but communicated transparently. The combination of belief intervals and natural language generation provides a concrete path to explain complex models with both quantitative rigor and semantic clarity, fostering trust, traceability, and human–machine collaboration in critical decision-making contexts.

Acknowledgements

The authors thank the reviewers for their valuable comments.

Author Contributions

JF conceived the study, developed the theoretical framework, and led the overall research design. NB contributed to the formalization of the Dempster–Shafer integration, provided critical revisions, and supervised the methodological design. MH and AA Conducted the experiments, and analyzed the results.

Statements and Declarations

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work received no external funding.

Data availability

All data and code are available at https://github.com/jfrez/DSExplainer.

References

[1] Molnar T. *Interpretable Machine Learning*. Leanpub, 2022.

[2] Murdoch WJ, Singh C, Kumbier K et al. Interpretable machine learning: Definitions, methods, and applications. *arXiv preprint arXiv:190104592* 2019; .

[3] Arrieta AB, Díaz-Rodríguez N, Ser JD et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* 2020; 58: 82–115.

[4] Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. pp. 4765–4774.

[5] Slack D, Hilgard S, Singh S et al. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In *Advances in Neural Information Processing Systems*, volume 34. pp. 9391–9404. NeurIPS 2021.

[6] Sensoy M, Kaplan L and Kandemir M. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. pp. 3183–3193.

[7] Watson DS, O'Hara J, Tax N et al. Explaining predictive uncertainty with information theoretic shapley values. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Tong Z, Xu P and Denoeux T. An evidential classifier based on dempster-shafer theory and deep learning. *Neurocomputing* 2021; 450: 275–293.

[9] Dempster AP. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics* 1967; 38(2): 325–339.

[10] Shafer G. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[11] Smets P and Kennes R. The transferable belief model. *Communications in Statistics – Theory and Methods* 1994; 23(5): 359–411.

[12] Slack D, Krishna S, Lakkaraju H et al. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence* 2023; 5: 873–883.

[13] Wang B, Li Y, Zhou J et al. Can llm assist in the evaluation of the quality of machine learning explanations? *arXiv preprint arXiv:250220635* 2024; .

[14] Bilal A, Ebert D and Lin B. LLMs for explainable AI: A comprehensive survey. *ACM Trans Intelligent Systems and Technology* ; .

[15] Shapley LS. A value for n-person games. *Contributions to the Theory of Games* 1953; 2(28): 307–317.

[16] Covert I, Lundberg SM and Lee S. Understanding global feature contributions via additive importance and interaction measures. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Includes Shapley interaction discussion; see also NeurIPS 2021 "Explaining by Removing".

[17] Hedström A, Weber L, Krakowczyk D et al. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research* 2023; 24(34): 1–11. URL https://www.jmlr.org/papers/v24/22-0142.html.

[18] Antorán J, Bhatt U, Adel T et al. Getting a CLUE: A method for explaining uncertainty estimates. In *Advances in Neural Information Processing Systems (NeurIPS)*. URL https://proceedings.neurips.cc/paper/2020/hash/5ca1536b6fcbf43d0f627f6b6f76a5f3-Abstract.html.

[19] Kendall A and Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*. Versiones previas: arXiv:1703.04977.

[20] Yager RR. On the dempster-shafer framework and new combination rules. *Information Sciences* 1987; 41(2): 93–137.

[21] Zhang ZY Zhen and Wang Y. An evidential classifier based on dempster-shafer theory and deep learning. *arXiv preprint arXiv:210313549* 2021; .

[22] Ribeiro MT, Singh S and Guestrin C. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. pp. 1135–1144.

[23] Sundararajan M, Taly A and Yan Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.