SAME PREDICTIONS, BETTER EXPLANATIONS: INTUITIONISTIC FUZZY HESITATION REDUCES SHAP INFIDELITY

CENGIZHAN TOPCU[1], SEVIL SENTURK[1], ASLI KAYA KARAKUTUK[1]

[1]Department of Statistics, Faculty of Science, Eskisehir Technical University, Eskisehir 26470, Turkey
E-MAIL: cengizhan@eskisehir.edu.tr, sdeligoz@eskisehir.edu.tr, aslikaya@eskisehir.edu.tr

Abstract:

Atanassov's intuitionistic fuzzy sets encode a hesitation degree, $\pi = 1 - \mu - \nu$, that quantifies how much evidence remains uncommitted in a decision. Despite its theoretical appeal, π has seen little use in applied settings, particularly in explainable AI. We test the idea head-on: an intuitionistic fuzzy classifier and a Type-1 counterpart are paired on the Pima Indians Diabetes data. Rule base, sigmoid membership functions, and cross-validation splits are all shared. Only π differs. A Mann–Whitney U test shows that misclassified cases carry markedly higher hesitation ($p < 10^{-85}$, rank-biserial $r = 0.43$). The two systems, by construction, classify every one of the 3 840 test instances identically, yet the IFS variant cuts SHAP infidelity by roughly a fifth (20.5 %) and LIME infidelity by a comparable margin (21.3 %). Because the class labels never change, the entire improvement lives in explanation quality. Taken together, the findings point to π as a low-cost reliability flag for clinical decision support: it signals when an explanation deserves closer inspection.

Keywords:

Intuitionistic fuzzy sets; Hesitation degree; Explainable AI; SHAP; Infidelity

1. Introduction

Clinicians increasingly rely on machine-learning classifiers for decision support, yet many remain sceptical, largely because the models offer no window into their reasoning. SHAP [1] and LIME [2] attempt to fill that gap by attributing each prediction to individual input features, something a clinician can actually examine. These attributions, though, turn out to be fragile themselves. Swap the random seed or redraw the perturbation sample and the feature rankings may rearrange considerably [3, 4]. An explanation that cannot be reproduced has little practical value in a clinical setting.

Fuzzy inference systems address this issue through a different approach. They encode knowledge in the form of linguistic rules that clinicians can read directly [5]. In standard Type-1 fuzzy sets, assigning a single membership value in [0, 1] to each input is functional up to a point. However, this formulation cannot distinguish between definite belonging and a complete lack of information about the input. Intuitionistic fuzzy sets (IFS), introduced by Atanassov (1986), separate these two situations [6]. In IFS, each element is assigned both a membership degree μ and a non-membership degree ν, subject to the constraint $\mu + \nu \leq 1$. The remaining portion, $\pi = 1 - \mu - \nu$, is called the hesitation degree and serves as a direct indicator of how insufficient the available evidence is. A high π value indicates that the available data do not support a reliable decision in either direction.

Within this framework, the main research question of the present study is as follows. Does retaining π during the inference process improve the fidelity of post-hoc explanations? The intuitive expectation is straightforward. In regions where the model is uncertain, the decision surface is flatter. The SHAP method proposed by Lundberg and Lee (2017) approximates this surface with a local linear model [1]. On a flatter surface, the linear approximation error is naturally smaller. If π captures exactly these uncertain regions, it has the potential to provide a reliability signal without any additional computational cost.

IFS-based classifiers have been applied in several areas of the literature. Ha´jek and Olej (2015) used them in credit scoring [7], Atanassov et al. (2023) in deep learning architectures [8], Ha´jek et al. (2021) in time-series forecasting [9], and Koc¸ak et al. (2023) in rule-based explainability [10]. The study most closely related to ours is that of Wang et al. (2023) [11], who developed an interpretable IFS stock-prediction model that ranks features by perturbing π. However, that ranking bypasses the Shapley framework entirely and therefore gives up its ax-

iomatic guarantees. On the SHAP side, Watson et al. (2023) extended Shapley values to predictive uncertainty through information-theoretic characteristic functions [12]. Their formulation, however, does not establish any connection with IFS. The present study aims to fill the gap between these two lines of research.

Two hypotheses guide the study. Both are tested under controlled conditions on the Pima Indians Diabetes dataset.

H1 (Hesitation as an uncertainty indicator). Misclassified instances exhibit statistically significantly higher π values than correctly classified ones. This finding indicates that the hesitation degree functions as a reliability signal.

H2 (Explanation fidelity improvement). Holding every other component constant, an IFS-based classifier achieves lower SHAP and LIME infidelity than its Type-1 counterpart. This improvement is traceable solely to π.

This study makes two contributions. First, it demonstrates empirically that the hesitation degree improves explanation fidelity. Second, it establishes this effect within a strict ceteris paribus design, eliminating the possibility that the observed improvement originates from any source other than π.

2. Background

2.1. Intuitionistic fuzzy sets

Formally, an intuitionistic fuzzy set A over a universe X is written

$$A = \{\langle x, \mu_A(x), \nu_A(x)\rangle \mid x \in X\}, \tag{1}$$

Here $\mu_A(x)$ is the membership degree and $\nu_A(x)$ the non-membership degree, subject to $\mu_A(x) + \nu_A(x) \le 1$ for every $x \in X$. The leftover margin is the hesitation degree:

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x). \tag{2}$$

When $\pi = 0$ everywhere, the IFS collapses to an ordinary fuzzy set. Any strictly positive π, however small, encodes a shade of indeterminacy that Type-1 logic has no room for.

Two IFS A and B intersect and unite under min/max as

$$A \cap B = \langle \min(\mu_A, \mu_B),\ \max(\nu_A, \nu_B)\rangle, \tag{3}$$

$$A \cup B = \langle \max(\mu_A, \mu_B),\ \min(\nu_A, \nu_B)\rangle. \tag{4}$$

The result that follows from (3) is that $\pi_{A\cap B}(x) \ge \pi_A(x)$ whenever $\nu_B(x) > \nu_A(x)$. That is, when antecedents are stacked, the indeterminacy can increase rather than decrease. No such behaviour exists in classical fuzzy logic.

2.2. Mamdani-type intuitionistic fuzzy inference

In a Mamdani-type IFS engine, knowledge is stored in rule form. Rule R_k is written as

$$R_k : \text{IF } x_1 \text{ is } \tilde{A}_{k1} \text{ AND } \dots \text{ AND } x_n \text{ is } \tilde{A}_{kn} \text{ THEN } y \text{ is } \tilde{B}_k, \tag{5}$$

where $\tilde{A}_{kj} = \langle \mu_{kj}, \nu_{kj}\rangle$ are intuitionistic fuzzy sets on each input x_j. Rule k fires with strength

$$w_k^{\mu} = \min_{j=1}^{n} \mu_{kj}(x_j), \qquad w_k^{\nu} = \max_{j=1}^{n} \nu_{kj}(x_j). \tag{6}$$

What remains after membership and non-membership have been accounted for, $\pi_k = 1 - w_k^{\mu} - w_k^{\nu}$ is the rule-level hesitation. Centroid defuzzification blends all K rules into a crisp output:

$$\hat{y} = \frac{\sum_{k=1}^{K} (w_k^{\mu} - w_k^{\nu}) \cdot c_k}{\sum_{k=1}^{K} (w_k^{\mu} - w_k^{\nu})}, \tag{7}$$

where c_k is the centroid of consequent $\tilde{B}_k$. The Type-1 variant sets $\nu_{kj} = 1 - \mu_{kj}$, forcing $\pi_k = 0$ and collapsing (7) to the standard Mamdani centroid. When hesitation is allowed to be positive, though, rules carrying large π_k weigh less. The factor $(w_k^{\mu} - w_k^{\nu})$ shrinks, and the system therefore discounts its own uncertain evidence. We suspect this built-in dampening is what connects IFS to better explanation fidelity, and Section 4 tests that suspicion directly.

2.3. SHAP and infidelity

SHAP assigns every feature i a value ϕ_i reflecting its marginal contribution to the prediction. Given a model f over feature set $N = \{1, \dots, n\}$ and a coalition $S \subseteq N \setminus \{i\}$:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left[\nu(S \cup \{i\}) - \nu(S)\right], \tag{8}$$

where $\nu(S) = \mathbb{E}[f(x) \mid x_S]$. KernelSHAP [1] turns (8) into a weighted least-squares problem:

$$\arg\min_{\phi} \sum_{z' \in Z} \pi_{\text{SHAP}}(z') \left[f(h_x(z')) - \phi_0 - \sum_{i=1}^{n} \phi_i z_i' \right]^2, \tag{9}$$

where $z' \in \{0, 1\}^n$ indicates active features, $h_x(z')$ maps coalitions to feature space, and $\pi_{\text{SHAP}}(z')$ is the Shapley kernel weight.

The degree to which an explanation faithfully tracks the actual behaviour of the model is a separate question. Yeh et al. [13] formalised this question with the infidelity metric:

$$\text{INFD}(\phi, f, x) = \mathbb{E}_e \left[(e^{\top}\phi - (f(x) - f(x - e)))^2 \right], \tag{10}$$

where e is a random perturbation. A low infidelity score means the linear approximation in ϕ tracks the model's true sensitivity closely. This is precisely the quantity we aim to measure.

3. Methodology

3.1. Ceteris paribus design

In this experiment, the fact that the two classifiers produce identical predictions was achieved by design and is the expected output. If everything other than π is held constant, any observed difference in explanation quality can only be attributed to π. Both systems were built from the same components. On each cross-validation fold, a shallow decision tree with depth four and minimum leaf size ten is trained. Every root-to-leaf path in this tree is converted into a fuzzy IF-THEN rule. The crisp split points are then softened with sigmoid functions: for the condition $x_j \leq \theta$, the membership value is computed as $\mu(x_j) = \sigma(k \cdot (\theta - x_j))$. The parameter k controls how sharp the transition is. The steepness value, the rule structure, the data partitioning, and the random seeds are all shared between the two systems. The only difference between them is whether π is retained.

On the IFS side, each condition additionally produces a non-membership degree $\nu(x_j) = \sigma(k \cdot (x_j - \theta))$, and the constraint $\mu + \nu \leq 1$ is satisfied by construction. Instance-level hesitation is computed from two components. The first is a *conflict* term, which measures how evenly the firing strengths are distributed across classes. The second is an *ignorance* term, which decays exponentially as the total amount of evidence increases. The relative contribution of these two components is controlled by λ_π and β_π. Class probability is derived from the Chen–Tan score function $S = \mu_{\text{out}} - \nu_{\text{out}}$, widely used in Atanassov set theory, and then rescaled to [0, 1]. In the Type-1 system, class probability is determined by the firing-strength-weighted average of consequent values. It can be shown algebraically that the outputs of the two systems coincide exactly when $\pi = 0$.

The comparison rests on a single algebraic property. Class 1 is assigned whenever the positive evidence w_m exceeds the negative evidence w_b. The factor $(1 - \pi)$ in the IFS score is always strictly positive. Because this factor cancels in the sign comparison, the two classifiers are necessarily forced to produce the same class label for every input. This agreement is not merely an empirical observation; it is a mathematically provable property. All differences between the two systems therefore emerge not in the class labels, but in the probability values. These probability values are precisely what SHAP and LIME interrogate.

3.2. Data and preprocessing

The Pima Indians Diabetes dataset created by Smith et al. (1988) was used in this study [14]. The dataset contains 768 instances, eight physiological measurements, and a binary label indicating diabetes onset within five years. To prevent data leakage, all preprocessing steps were carried out separately within each cross-validation fold. Glucose, blood pressure, skin thickness, insulin, and body mass index include zero entries that are physiologically impossible. These entries were treated as missing values and imputed with the median computed from the training subset of the corresponding fold. The seven features with the highest information content were selected using mutual information, after which min–max scaling to [0, 1] was applied using training data only. There is a specific reason for choosing the Pima dataset. Its baseline accuracy of approximately 76 % produces a sufficient number of misclassifications for testing H1.

3.3. Evaluation protocol

Repeated stratified five-fold cross-validation was applied five times, yielding 25 folds and a total of 3 840 test evaluations. Before the main experiment, a grid search was conducted over four IFS-specific parameters (k, δ, λ_π, β_π). The optimisation criterion was the expected AURC under the product confidence strategy. The best-performing parameter combination was $k = 20$, $\delta = 0$, $\lambda_\pi = 0.2$, and $\beta_\pi = 2.0$. All four parameters concern only the hesitation mechanism. The Type-1 system inherits the optimised k value, but because it has no hesitation-related parameter, the ceteris paribus condition holds automatically.

For H1, the π values corresponding to correct and incorrect predictions were compared using the Mann–Whitney U test, and effect size was quantified by the rank-biserial correlation. H2 required stricter experimental control. Both systems derived their SHAP and LIME explanations from a shared full-data model. The training set, the 100 k-means background centroids, the perturbation count, and the random seed were kept identical across both systems. Infidelity was computed instance by instance according to equation (10) proposed by Yeh et al. (2019) [13]. For each instance, 50 perturbation draws were used with a noise radius of 0.15.

Given an input x, instance-level hesitation is computed as the weighted average of the rule-level values:

$$\bar{\pi}(x) = \frac{\sum_{k=1}^{K} w_k^{\mu} \cdot \pi_k(x)}{\sum_{k=1}^{K} w_k^{\mu}}, \tag{11}$$

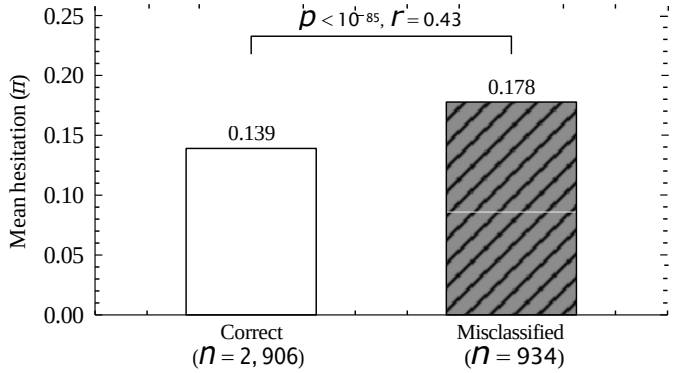

FIGURE 1. Mean hesitation degree (π) for correctly classified and misclassified instances across all 25 test folds. Brackets: Mann–Whitney U p-value and rank-biserial r.

where $\pi_k(x) = 1 - w_k^{\mu}(x) - w^{v}(x)$ denotes the hesitation value of the kth rule. The use of w_k^{μ} as the weight prevents rules with low firing strength from exerting a disproportionate influence on the overall estimate.

4. Results

4.1. Hesitation degree as a reliability signal

In this subsection, we examine whether the hesitation degree is associated with classification outcome. To this end, the hesitation levels of correctly and incorrectly classified instances are compared, and H1 is evaluated within this framework.

Figure 1 visually summarises the H1 results. The mean hesitation value of the misclassified instances was $\bar{\pi} = 0.178$, whereas the corresponding value for the correctly classified instances was $\bar{\pi} = 0.139$. Table 1 presents the detailed distributional breakdown.

TABLE 1. Hesitation degree (π) by classification outcome.

	Correct	Misclassified	Statistic
Mean π	0.139	0.178	$U = 1.93 \times 10^6$
n	2 906	934	$p < 10^{-85}$
Effect size	rank-biserial $r = 0.43$		

A Mann–Whitney U test rejected the null hypothesis at $p < 10^{-85}$ (rank-biserial $r = 0.43$). This result means that, in a randomly selected correct–incorrect pair, the probability that the misclassified instance carries a higher π value is approximately 70 %. Although the absolute difference between the means appears small at 0.039, this difference was observed consistently across all 25 folds without exception. The Spearman correlation between π and the binary error indicator was $\rho = 0.316$ ($p < 10^{-89}$), confirming the monotonic association. The practical implication is that a prediction with a high π value may be used as an early warning indicator of unreliability even before any explanation is computed. For example, predictions whose π value exceeds a predetermined threshold may be automatically flagged for the clinician.

4.2. Explanation fidelity gains under identical predictions

In this subsection, we investigate whether two systems that yield identical class decisions nevertheless differ in terms of explanation fidelity. In this way, the analysis isolates whether the observed differences arise from the hesitation mechanism itself rather than from any change in predictive decisions.

Table 2 presents the main finding of the study. As shown in Section 3.1, the sign comparison that determines the class decision is independent of π. Consistent with this mathematical derivation, the two systems produced the same class label for all 3 840 test instances. Nevertheless, the IFS variant reduced SHAP infidelity by 20.5 % and LIME infidelity by 21.3 %.

TABLE 2. Explanation fidelity when only π differs; prediction agreement is perfect across 3 840 instances.

Metric	IFS	Type-1	Δ
SHAP infidelity ↓	**0.075**	0.094	−20.5 %
LIME infidelity ↓	**0.074**	0.094	−21.3 %
Jaccard stability ↑	**0.731**	0.717	+2.0 %
SHAP sensitivity ↓	**0.397**	0.457	−13.1 %
Classification agreement: 3 840 / 3 840 (100 %)			

This improvement does not arise from different predictions. The predictions are exactly identical. The only element that differs is how faithfully these predictions are explained by SHAP and LIME. The source of the observed improvement is π. The mechanism can be explained through equation (7). Rules with high π_k values receive lower weights through the factor $(w_k^{\mu} - w^{v})$. This means that uncertain evidence is suppressed during the inference process. The regions in which SHAP's linear approximation produces its largest errors are precisely these high-uncertainty regions. The weighting mechanism introduced by IFS smooths the decision surface in those regions and therefore reduces linear approximation error. The simultaneous improvement in both SHAP and LIME infidelity indicates that this gain is not a property of a specific explanation method, but rather of the model structure itself.

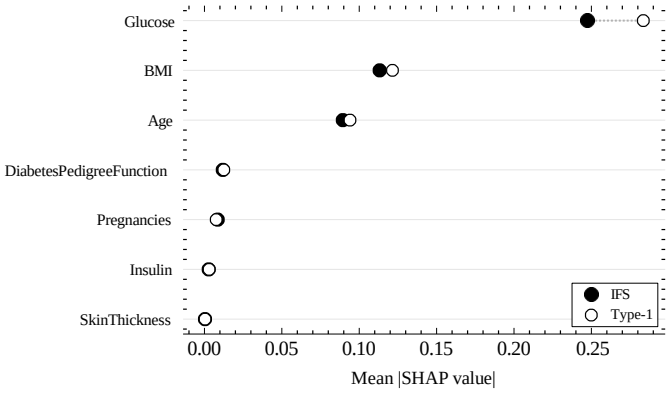

FIGURE 2. Mean |SHAP| values for each feature (Cleveland dot plot). Filled circles: IFS; open circles: Type-1. Rankings are identical across both systems.

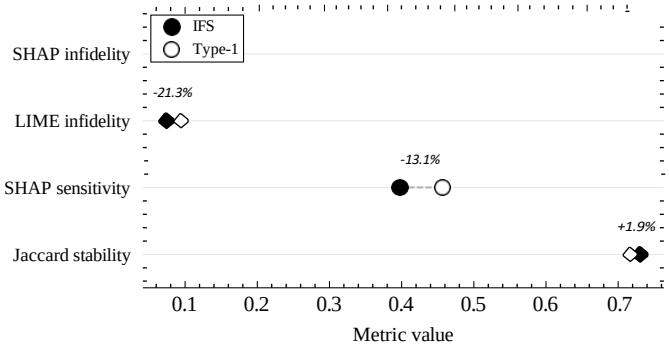

FIGURE 3. Explanation quality metrics under ceteris paribus conditions. Filled circles: IFS; open circles: Type-1. Percentage deltas are shown above each connector. Lower is better for infidelity and sensitivity; higher is better for stability.

4.3. Feature importance concordance

To show that the improvement in explanation fidelity is not driven by a change in feature-importance ordering, the mean |SHAP| values obtained from the two systems are compared separately in this subsection.

The mean |SHAP| values displayed as a Cleveland dot plot in Figure 2 reveal identical feature rankings in both systems. Glucose ranks first, followed by BMI and then Age. This is fully consistent with clinical expectations [14].

In the Cleveland dot plot shown in Figure 2, the IFS point (filled circle) for each feature is located either to the left of or very close to the Type-1 point (open circle). This observation indicates that the fidelity gain documented under H2 does not arise from a reordering of feature importance. The same features remain important in both systems. Once π is introduced, the only aspect that changes is the fidelity with which the attributions assigned to those features reflect the model's actual behaviour.

5. A Geometric Account of Fidelity Improvement

5.1. Connecting H1 and H2

In this subsection, the findings for H1 and H2 are considered jointly in order to discuss, in an integrated manner, how the hesitation degree influences both the decision surface and explanation fidelity.

The findings obtained from H1 and H2 support one another. H1 showed that π is statistically associated with predictive uncertainty (Figure 1). H2 demonstrated that this relationship translates into a measurable improvement in explanation fidelity (Figure 3). Both results were obtained without changing a single class label. The proposed mechanism is geometric in nature. Equation (7) assigns lower weight to rules with high hesitation. This operation smooths the decision surface in uncertain regions. Since SHAP fits a local hyperplane, a smoother target surface implies a smaller approximation error and therefore lower infidelity.

One situation in which π falls short should nevertheless be noted. In the regions where low scores were observed, the Spearman correlation between π and LIME Jaccard stability was only $\rho = 0.187$ ($p = 0.19$, $n = 50$ XAI evaluation instances). Limiting the sample size to fifty was not an arbitrary choice. Computing Jaccard stability required repeated LIME runs for each instance, and within our shared-resource ceteris paribus design, this made a larger evaluation practically infeasible. Fifty instances therefore represented the upper limit of what could be implemented. For this reason, a Type II error remains a genuine possibility.

Even so, the absence of statistical significance is not theoretically surprising. Infidelity measures how closely a linear approximation follows the model's local gradient, and this is a property of decision-surface geometry. Jaccard stability, by contrast, reflects how reproducibly LIME selects the same features across independent perturbation draws. This is a property of sampling variance. The geometric smoothing introduced by π through equation (7) should, by its nature, improve the former. However, it does not provide a clear mechanism for reducing the stochastic noise involved in the latter.

The WDBC result ($\rho = -0.505$, $p = 1.8 \times 10^{-4}$) provides a useful counterpoint to this interpretation. In a cleaner problem with a higher signal-to-noise ratio, the consistency-related

effect of π becomes visible. In Pima, by contrast, the decision boundary is inherently noisier. The effect may still be present, but it remains below the level that can be detected reliably with fifty instances.

5.2. A note on selective classification

We also checked whether π could help with selective classification, where a model abstains on low-confidence inputs. A paired t-test on the hesitation-based strategy returned $t = 2.96$, $p = 0.007$ over 25 folds. This result is statistically significant, but the practical magnitude is negligible (E-AURC difference on the order of 10^{-3}). The product confidence strategy did not even reach significance ($p = 0.13$).

This pattern may be regarded as informative rather than disappointing. Selective classification asks whether the model should refrain from borderline predictions. Explanation fidelity asks whether the predictions that are made can be summarised accurately. These are different questions. By construction, the hesitation degree suppresses uncertain rules in the defuzzified output (equation 7) without changing the sign of the output. For this reason, it is reasonable to expect π to reshape the decision surface, making it smoother and easier to approximate, without substantially altering which instances fall above or below the classification threshold. Rather than undermining the geometric mechanism proposed in Section 5.1, the selective classification results are consistent with it. π acts on the surface, not on the boundary.

5.3. Corroboration from WDBC

Does this pattern generalise? To examine this question, the same preprocessing pipeline and parameter settings from Pima were applied to the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [15]. This is a considerably easier problem, with accuracy above 94 %. Despite the much smaller number of errors, the hesitation signal remained intact: $\pi_{\text{incorrect}} = 0.402$ versus $\pi_{\text{correct}} = 0.288$ (Mann–Whitney $p = 1.5 \times 10^{-72}$). On WDBC, the Spearman correlation between π and Jaccard stability was $\rho = -0.505$ ($p = 1.8 \times 10^{-4}$). Here it is significant where Pima fell short. Dataset characteristics therefore appear to modulate how far the explanatory signal of π extends. On WDBC, SHAP infidelity decreased by 46.3 % (IFS: 0.073, Type-1: 0.135), approximately twice the reduction observed on Pima. We reserve the full H2-equivalent analysis on WDBC for a longer companion paper.

5.4. Limitations and future directions

This design has clear limitations that should be stated explicitly. Neither system was tuned for raw predictive accuracy. The grid search targeted only the hesitation parameters, while all shared components were left unchanged. A jointly optimised IFS could plausibly improve both predictive accuracy and explanation quality. However, disentangling the contribution of π from the benefits of tuning would require a separate investigation, which we plan to pursue in future work. In practice, π could be displayed alongside SHAP waterfall plots in a clinical dashboard, providing clinicians with a quick visual cue about explanation reliability. No modification to the existing SHAP infrastructure would be required. A more ambitious line of work would integrate π directly into the Shapley-value framework by defining a hesitation-aware characteristic function. The empirical evidence presented here provides the necessary foundation for such a step.

6. Conclusion

We compared an intuitionistic fuzzy inference system with its Type-1 counterpart on the Pima diabetes dataset while holding every component constant except the hesitation degree. Misclassified instances carried significantly higher π values, confirming H1. The mathematical structure guarantees identical predictions, and all 3 840 test instances confirmed this property. Nevertheless, the IFS reduced SHAP infidelity by 20.5 % and LIME infidelity by 21.3 %, confirming H2.

The improvement observed on the WDBC dataset was even more pronounced than on Pima. SHAP infidelity decreased by 46.3 %. The correlation between π and stability, which did not reach statistical significance on Pima, was significant on WDBC. Overall, π does not alter the predictions produced by the model. However, it makes the explanations of those predictions substantially more faithful. This property represents a worthwhile gain for clinical decision-support systems.

References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4766–4777, 2017.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD*, pp. 1135–1144, 2016.

[3] D. Alvarez-Melis and T. S. Jaakkola, “On the robustness of interpretability methods,” in *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

[4] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, “Do feature attribution methods correctly attribute features?,” in *Proc. AAAI*, vol. 36, pp. 9623–9633, 2022.

[5] D. D. Nauck and R. Kruse, “Obtaining interpretable fuzzy classification rules from medical data,” *Artif. Intell. Med.*, vol. 16, no. 2, pp. 149–169, 1999.

[6] K. T. Atanassov, “Intuitionistic fuzzy sets,” *Fuzzy Sets Syst.*, vol. 20, no. 1, pp. 87–96, 1986.

[7] P. Ha´jek and V. Olej, “Intuitionistic fuzzy neural network: The case of credit scoring using text information,” in *Proc. EANN*, CCIS 517, pp. 337–346, 2015.

[8] K. Atanassov, S. Sotirov, and T. Pencheva, “Intuitionistic fuzzy deep neural network,” *Mathematics*, vol. 11, no. 3, p. 716, 2023.

[9] P. Ha´jek, V. Olej, W. Froelich, and J. Novotny, “Intuitionistic fuzzy neural network for time series forecasting,” in *Proc. AIAI*, IFIP AICT 627, pp. 411–422, 2021.

[10] C. Kocak, E. Egrioglu, and E. Bas, “A new explainable robust high-order intuitionistic fuzzy time-series method,” *Soft Comput.*, vol. 27, pp. 1783–1796, 2023.

[11] W. Wang, W. Lin, Y. Wen, X. Lai, P. Peng, Y. Zhang, and K. Li, “An interpretable intuitionistic fuzzy inference model for stock prediction,” *Expert Syst. Appl.*, vol. 213, p. 118908, 2023.

[12] D. S. Watson, J. O’Hara, N. Tax, R. Mudd, and I. Guy, “Explaining predictive uncertainty with information theoretic Shapley values,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[13] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, “On the (in)fidelity and sensitivity of explanations,” in *Advances in Neural Information Processing Systems*, vol. 32, pp. 10965–10976, 2019.

[14] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” in *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 261–265, 1988.

[15] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear feature extraction for breast tumor diagnosis,” in *Proc. IS&T/SPIE Electron. Imaging*, vol. 1905, pp. 861–870, 1993.