applied sciences

MDPI

Article

An Explainable Fuzzy Multi-Criteria Decision-Making Framework with SHAP-Guided Rule Extraction for Transparent Decision Support Under Uncertainty

Jesús Alberto Rodríguez-Flores [1], Alexander Sánchez-Rodríguez [2,*], Yandi Fernández-Ochoa [1], Gelmar García-Vidal [2], Alexis Cordovés-García [1] and Reyner Pérez-Campdesuñer [2]

[1] Engineering Systems Research and Consulting Group (GICSI), Universidad UTE, Quito 170527, Ecuador; jesus.rodriguez@ute.edu.ec (J.A.R.-F.); yandi.fernandez@ute.edu.ec (Y.F.-O.); alexis.cordoves@ute.edu.ec (A.C.-G.)

[2] Business Administration Research Group (GADEP), Universidad UTE, Quito 170527, Ecuador; gelmar.garcia@ute.edu.ec (G.G.-V.); reyner.perez@ute.edu.ec (R.P.-C.)

* Correspondence: alexander.sanchez@ute.edu.ec; Tel.: +593-987171408

Featured Application

The proposed framework can support supplier selection in uncertain industrial and supply chain environments by producing transparent, interpretable, and auditable rankings. By combining FAHP–FTOPSIS ranking with surrogate-based SHAP analysis and linguistic IF–THEN rule extraction, it enables decision-makers to trace criterion-level contributions, understand trade-offs, and justify supplier-selection outcomes.

Abstract

Conventional fuzzy multi-criteria decision-making (MCDM) methods support ranking under uncertainty but often provide limited explanation of why alternatives are preferred. This study proposes an explainable fuzzy decision-making framework that integrates the Fuzzy Analytic Hierarchy Process (FAHP) and Fuzzy TOPSIS with surrogate modeling, SHAP-based analysis, and linguistic rule extraction. The main contribution is an explanation layer that preserves the original FAHP–FTOPSIS ranking structure while decomposing ranking scores into criterion-level contributions and transforming recurrent attribution patterns into IF–THEN rules. The framework is evaluated through a supplier-selection case study using expert fuzzy evaluations, local perturbation analysis, leave-one-supplier-out cross-validation, and a synthetic benchmark. The results show that the fuzzy MCDM layer produces discriminative rankings and that the top-ranked supplier remains comparatively stable under perturbations. Among the tested surrogates, the Random Forest Regressor achieved the strongest local fidelity, outperforming linear regression and a shallow decision tree. SHAP analysis showed ordinal alignment between FAHP weights and global criterion importance, while the extracted rules achieved high coverage, consistency, and threshold stability. The framework is useful for researchers, decision analysts, procurement managers, and supply chain professionals who require transparent, interpretable, and auditable multicriteria decisions under uncertainty.

Keywords: fuzzy systems; multicriteria decision making; decision support systems; rule extraction; interpretability; SHAP

Academic Editor: Diana Kalibatiene

Received: 13 April 2026
Revised: 14 May 2026
Accepted: 14 May 2026
Published: 21 May 2026

1. Introduction

Fuzzy systems provide a mathematically grounded framework for representing and aggregating imprecise, linguistic, and subjective information in complex decision-making environments. In organizational, industrial, and engineering contexts, problems such as supplier selection, project prioritization, risk assessment, and resource allocation frequently involve uncertainty, ambiguity, and multiple conflicting criteria. In this setting, Multi-Criteria Decision-Making (MCDM) methods have become a well-established family of analytical tools for structuring decision problems and generating preference rankings by aggregating heterogeneous criteria within a coherent evaluation framework [1,2].

However, conventional MCDM approaches often rely on crisp numerical inputs and deterministic preference structures, which may be restrictive when expert judgments are expressed through linguistic terms such as "high", "moderate", or "low". Fuzzy set theory addresses this limitation by formally representing vagueness and partial membership, allowing decision-makers to preserve the semantic richness of expert evaluations without forcing artificial numerical precision [3]. Within this tradition, Fuzzy Analytic Hierarchy Process (Fuzzy AHP) and Fuzzy Technique for Order Preference by Similarity to Ideal Solution (Fuzzy TOPSIS) are among the most widely used methods for multicriteria problems under uncertainty. Fuzzy AHP extends pairwise comparison logic to fuzzy environments for deriving criteria weights from imprecise judgments [4,5], whereas Fuzzy TOPSIS ranks alternatives according to their relative closeness to fuzzy ideal solutions [6]. Systematic reviews confirm the consolidation of fuzzy MCDM methods, including hybrid approaches, consistency improvements, and more expressive representations of uncertainty [7,8].

Despite these advances, fuzzy MCDM models often remain limited in their ability to explain the internal logic behind ranking outcomes. In most applications, the process yields weights, scores, or rankings, but does not explicitly show why an alternative is preferred, which criteria drive the result, or how trade-offs among criteria shape the final decision. This limitation is particularly relevant in contexts such as supplier evaluation, where decisions must be not only analytically robust but also traceable, communicable, and auditable. Consequently, the value of MCDM models increasingly depends on their ability to provide intelligible and explainable decision logic, in line with the growing emphasis on explainable artificial intelligence in decision-support systems [2,9].

This need has motivated the incorporation of interpretability-oriented approaches, particularly from Explainable Artificial Intelligence (XAI). XAI seeks to clarify how analytical models transform inputs into outputs, supporting trust, accountability, and responsible use in human-centered decision contexts [9,10]. Among existing XAI techniques, SHAP (SHapley Additive exPlanations) has gained relevance because of its foundation in cooperative game theory and its capacity to provide consistent local and global feature-attribution explanations. By decomposing model outputs into additive contributions, SHAP allows criterion-level effects to be analyzed without modifying the original decision structure [11].

Nevertheless, the integration between fuzzy MCDM and explainability techniques remains limited and fragmented. Existing approaches often apply interpretability tools as post hoc additions without ensuring sufficient coherence between the fuzzy decision model and its explanation layer. As a result, explanations may remain numerical or visual, but may not provide a compact, human-readable representation of the decision logic. This gap highlights the need for unified frameworks in which fuzzy evaluation, ranking mechanisms, surrogate approximation, and explanation models are systematically connected.

Motivated by this limitation, this study proposes an explainable fuzzy multi-criteria decision-making framework that integrates Fuzzy AHP, Fuzzy TOPSIS, surrogate modeling, SHAP-based analysis, and linguistic rule extraction within a unified decision architecture. The proposed approach first generates rankings under uncertainty and then approximates

the resulting fuzzy decision operator through a surrogate model. SHAP values are used to decompose ranking scores into criterion-level contributions, which are subsequently transformed into interpretable IF–THEN rules. In this way, the framework bridges the gap between numerical fuzzy-ranking outputs and symbolic decision explanations that are suitable for interpretation and auditability [11–13].

Unlike existing explainable MCDM and hybrid fuzzy–XAI approaches that mainly use attribution methods to visualize or quantify the influence of criteria on model outputs, the proposed framework advances a more structured explanation mechanism. Its novelty lies in explicitly connecting the FAHP–FTOPSIS ranking operator with surrogate-based SHAP decomposition and linguistic rule extraction. Thus, SHAP values are not used only as post hoc numerical explanations; they are transformed into compact IF–THEN rules that summarize recurrent contribution patterns while preserving the original fuzzy decision structure.

Accordingly, the general objective of this study is to develop and evaluate an explainable fuzzy multi-criteria decision-making framework that generates transparent and interpretable rankings under uncertainty. More specifically, this study pursues four objectives: first, to integrate FAHP and FTOPSIS into a coherent fuzzy decision operator for ranking alternatives based on linguistic expert evaluations; second, to approximate this operator through a surrogate model with sufficient local fidelity; third, to use SHAP values to decompose ranking outcomes into criterion-level contributions; and fourth, to transform these contribution patterns into linguistic IF–THEN rules that support transparent decision justification.

From a methodological standpoint, this paper's contribution is threefold. First, it formulates a unified decision architecture in which fuzzy weighting, fuzzy ranking, and explainability are treated as connected components rather than isolated procedures. Second, it extends the interpretability of fuzzy multicriteria ranking beyond feature-attribution analysis by generating linguistic rules derived from SHAP-based contribution structures. Third, it evaluates the explanatory reliability of the proposed framework by assessing surrogate fidelity, consistency between FAHP-derived weights and SHAP-based attributions, linguistic-rule interpretability, and local robustness to bounded input perturbations.

These contributions are evaluated through a reproducible supplier-selection case study complemented by controlled local perturbation analysis. The case study is used to instantiate the proposed framework and to illustrate its operation in a realistic multicriteria decision context. Given the limited number of original alternatives, the empirical section is not intended to provide population-level statistical generalization. Instead, the augmented perturbation-based dataset is used to assess local surrogate fidelity, SHAP-based explanation consistency, rule interpretability, and ranking robustness within the bounded decision space defined by the case study.

The remainder of this paper is organized as follows: Section 2 presents the theoretical background on fuzzy sets, Fuzzy AHP, Fuzzy TOPSIS, the fuzzy MCDM decision operator, and SHAP-based explainability. Section 2.7 reviews related work and identifies the research gap addressed by the proposed framework. Section 3 describes the materials and methods, including the architecture of the proposed framework, fuzzy aggregation, surrogate modeling, local perturbation, SHAP-based rule extraction, and the complete algorithmic workflow. Section 4 reports the empirical results of the supplier-selection case study, including fuzzy MCDM rankings, surrogate-model fidelity, SHAP explanations, rule extraction, sensitivity analysis, and robustness analysis. Section 5 discusses the theoretical, methodological, and practical implications of the findings, together with the main limitations and future research directions. Finally, Section 6 presents the conclusions of this study.

2. Theoretical Background

2.1. Fuzzy Sets and Triangular Fuzzy Numbers

Classical set theory assumes a binary membership structure in which an element either belongs to a set or does not belong to it. Although this representation is suitable for precise and well-defined phenomena, it is often insufficient for real-world decision problems involving ambiguity, vagueness, and partial truth. In multicriteria settings, expert judgments are frequently expressed through linguistic assessments rather than exact numerical values, reflecting subjective perceptions and incomplete information. Fuzzy set theory, introduced by Zadeh [3], provides a mathematically tractable way to represent this type of uncertainty by allowing elements to belong to a set with varying degrees of membership.

Formally, a fuzzy set $\widetilde{A}$ defined on a universe of discourse X is characterized by a membership function:

$$\mu_{\widetilde{A}} : X \to [0, 1]$$

where $\mu_{\widetilde{A}}(x)$ denotes the degree to which the element $x \in X$ belongs to the fuzzy set $\widetilde{A}$. A membership value of 0 indicates complete non-membership, a value of 1 indicates full membership, and intermediate values represent partial membership.

Among the different representations of fuzzy numbers, triangular fuzzy numbers (TFNs) are especially suitable for multicriteria decision-making because of their simplicity, intuitive interpretation, and computational efficiency. A triangular fuzzy number is defined as follows:

$$\widetilde{a} = (l, m, u)$$

where l, m, and u denote the lower, modal, and upper bounds, respectively, with:

$$l \leq m \leq u$$

Its membership function is given by:

$$\mu_{\widetilde{a}}(x) = \begin{cases} 0, & x < l, \\ \frac{x-l}{m-l}, & l \leq x \leq m, \\ \frac{u-x}{u-m}, & m \leq x \leq u, \\ 0, & x > u. \end{cases}$$

This representation assumes that m is the most plausible value, while l and u define the lower and upper limits of plausible variation. For this reason, TFNs are commonly used to encode linguistic terms such as “very low”, “low”, “medium”, “high”, or “very high” in fuzzy decision environments. Their use is widespread in fuzzy MCDM models because they offer a practical balance between expressive capacity and analytical manageability, particularly in foundational fuzzy AHP and fuzzy TOPSIS formulations.

In the proposed framework, expert evaluations are expressed linguistically and then transformed into triangular fuzzy numbers to preserve the uncertainty inherent in human judgment. Let:

$$A = \{A_1, \ldots, A_m\}$$

denote the set of alternatives and:

$$C = \{C_1, \ldots, C_n\}$$

the set of decision criteria. The fuzzy evaluation of alternative A_i with respect to criterion C_j is represented as follows:

$$\widetilde{x}_{ij} = (l_{ij}, m_{ij}, u_{ij}),$$

where l_{ij}, m_{ij}, and u_{ij} indicate the lower, modal, and upper bounds of the corresponding assessment. Collecting these evaluations for all alternatives and criteria yields the fuzzy decision matrix:

$$\widetilde{X} = [\widetilde{x}_{ij}]_{m \times n}$$

This matrix constitutes the basic representation of the decision problem under uncertainty. It transforms qualitative expert judgments into a structured mathematical form that preserves linguistic imprecision while enabling the formal operations required for Fuzzy AHP, Fuzzy TOPSIS, and the subsequent explainability layer.

2.2. Fuzzy Analytic Hierarchy Process

The Analytic Hierarchy Process (AHP), originally introduced by Saaty [14], is one of the most widely used methods for deriving criteria weights in multicriteria decision-making. Its central idea is to decompose a decision problem into a hierarchical structure and determine the relative importance of criteria through pairwise comparisons. In its classical form, however, AHP assumes that decision-makers can express their judgments using precise numerical values. This assumption may be restrictive in real-world contexts where expert assessments are uncertain, vague, or linguistically formulated rather than strictly quantitative.

To address this limitation, fuzzy extensions of AHP replace crisp comparison values with fuzzy numbers. This allows the method to capture the ambiguity of human judgments while preserving the comparative logic of the original AHP structure. Among the earliest and most influential formulations, Buckley [4] proposed a fuzzy hierarchical analysis based on triangular fuzzy numbers, which has become a standard reference in fuzzy weighting procedures.

Let:

$$\widetilde{A} = [\widetilde{a}_{ij}]_{n \times n}$$

denote the fuzzy pairwise comparison matrix, where $\widetilde{a}_{ij}$ expresses the fuzzy importance of criterion C_i relative to criterion C_j. In the present framework, each comparison value is represented by a triangular fuzzy number:

$$\widetilde{a}_{ij} = (l_{ij}, m_{ij}, u_{ij})$$

where l_{ij}, m_{ij}, and u_{ij} denote the lower, modal, and upper bounds of the expert judgment. To maintain reciprocal consistency, the following property is assumed:

$$\widetilde{a}_{ji} = \left(\frac{1}{u_{ij}}, \frac{1}{m_{ij}}, \frac{1}{l_{ij}} \right)$$

This formulation preserves the relative meaning of pairwise comparisons while incorporating uncertainty in a mathematically consistent way.

To derive the criteria weights, this study adopts the fuzzy geometric mean method proposed by Buckley [4]. For each criterion C_i, the fuzzy geometric mean is computed as follows:

$$\widetilde{g}_i = \left(\prod_{j=1}^{n} \widetilde{a}_{ij} \right)^{1/n}, i = 1, \ldots, n$$

The corresponding normalized fuzzy weight is then obtained as follows:

$$\widetilde{w}_i = \frac{\widetilde{g}_i}{\sum_{k=1}^{n} \widetilde{g}_k}, i = 1, \ldots, n$$

These fuzzy weights represent the relative importance of the criteria under uncertain pairwise judgments. When a crisp representation is required for subsequent computational steps, each fuzzy weight $\widetilde{w}_i = (l_i, m_i, u_i)$ is defuzzified using the centroid method:

$$w_i = \frac{l_i + m_i + u_i}{3}$$

After defuzzification, the resulting weights are normalized so that:

$$\sum_{i=1}^{n} w_i = 1$$

The final weights quantify the relative contribution of each criterion within the decision structure and are incorporated into the Fuzzy TOPSIS procedure for ranking the alternatives. Thus, Fuzzy AHP provides a weighting mechanism that preserves uncertainty in expert judgments while producing operational criteria weights for the multicriteria evaluation process.

2.3. Fuzzy TOPSIS

Building on the fuzzy evaluations and the criteria weights derived from Fuzzy AHP, the ranking stage is performed using Fuzzy TOPSIS. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), originally proposed by Hwang and Yoon [15], ranks alternatives according to their relative distances from positive and negative ideal solutions. Chen [6] extended this logic to fuzzy environments by allowing evaluations and weights to be represented as fuzzy numbers.

Given the fuzzy decision matrix $\widetilde{X}$ and the criteria weights w_j, the weighted fuzzy decision matrix is constructed as follows:

$$\widetilde{v}_{ij} = \widetilde{x}_{ij} \times w_j$$

The fuzzy positive ideal solution (FPIS) and fuzzy negative ideal solution (FNIS) are defined as follows:

$$A^* = \left(\widetilde{v}_1^*, \ldots, \widetilde{v}_n^*\right)$$

$$A^- = \left(\widetilde{v}_1^-, \ldots, \widetilde{v}_n^-\right)$$

where, for benefit criteria, $\widetilde{v}_j^* = \max_i \widetilde{v}_{ij}$ and $\widetilde{v}_j^- = \min_i \widetilde{v}_{ij}$, whereas for cost criteria this definition is reversed.

The distance between two triangular fuzzy numbers $\widetilde{a} = (l_a, m_a, u_a)$ and $\widetilde{b} = (l_b, m_b, u_b)$ is computed using the vertex method:

$$d(\widetilde{a}, \widetilde{b}) = \sqrt{\frac{(l_a - l_b)^2 + (m_a - m_b)^2 + (u_a - u_b)^2}{3}}$$

For each alternative A_i, the distances to the ideal solutions are calculated as follows:

$$D_i^* = \sum_{j=1}^{n} d(\widetilde{v}_{ij}, \widetilde{v}_j^*)$$

$$D_i^- = \sum_{j=1}^{n} d(\tilde{v}_{ij}, \tilde{v}_j^-)$$

The closeness coefficient is then obtained as follows:

$$CC_i = \frac{D_i^-}{D_i^* + D_i^-}$$

Alternatives are ranked in descending order of CC_i. Therefore, the closeness coefficient can be interpreted as a normalized decision score that reflects the relative desirability of each alternative within the fuzzy evaluation space.

2.4. The Fuzzy MCDM Decision Operator

The combination of Fuzzy AHP and Fuzzy TOPSIS defines an implicit decision operator that maps fuzzy evaluations into scalar ranking scores. Formally, let X_i denote the criterion vector associated with alternative A_i, $\tilde{X}$ the fuzzy decision matrix, and w the vector of criteria weights derived from Fuzzy AHP. The fuzzy multicriteria decision-making process can be represented as follows:

$$f : X_i \rightarrow CC_i$$

where CC_i is the closeness coefficient obtained through the Fuzzy TOPSIS procedure.

This formulation highlights that the FAHP–FTOPSIS pipeline operates as a deterministic transformation from uncertain, linguistically expressed inputs into normalized decision scores. Although each stage is analytically defined, the resulting mapping is not directly interpretable in terms of individual criterion contributions. This motivates the use of surrogate modeling and SHAP-based decomposition in the subsequent explainability stage.

2.5. Explainable Artificial Intelligence and SHAP

Explainable Artificial Intelligence (XAI) seeks to improve the transparency and interpretability of algorithmic decision systems by clarifying how model inputs influence outputs. As predictive and decision-support models become increasingly complex, explainability has become essential for ensuring trust, accountability, and responsible use of artificial intelligence, particularly when algorithmic recommendations support human decision-making [9,10].

Recent research shows that explainability is relevant not only in predictive modeling but also in broader decision-support contexts. For example, Lu et al. [16] developed an NLP-based approach for decoding urban policies through concise explanations, showing how computational models can make complex policy texts more accessible and interpretable for decision-makers. Although their work is situated in urban analytics rather than fuzzy MCDM, it reinforces the broader need for explanation mechanisms that translate complex analytical processes into concise and human-readable outputs.

Among existing XAI approaches, SHAP (SHapley Additive exPlanations), introduced by Lundberg et al. [11], is one of the most influential and theoretically grounded methods for feature-attribution analysis. SHAP is based on Shapley values from cooperative game theory, where the prediction task is interpreted as a game and each input feature is treated as a player contributing to the final outcome. The contribution of each feature is quantified by averaging its marginal effect over all possible feature coalitions, resulting in an attribution measure with strong theoretical justification.

Given a predictive model $f(x)$, SHAP represents the prediction in additive form as follows:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i$$

where ϕ_i denotes the contribution of feature i to the model output, and ϕ_0 denotes the baseline prediction, typically associated with the expected model output over a reference distribution.

The Shapley value associated with feature i is computed as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} [f(S \cup \{i\}) - f(S)]$$

where $N = \{1, \ldots, n\}$ is the full set of features and S denotes a subset of features that does not contain feature i. This expression measures the marginal contribution of feature i across all possible subsets S, weighted according to the Shapley combinatorial coefficients. Thus, SHAP provides a principled decomposition of model predictions into feature-level effects.

A key strength of SHAP is that it satisfies desirable properties such as local accuracy, missingness, and consistency. Local accuracy ensures that the sum of all feature contributions equals the model prediction relative to the baseline. Missingness assigns zero contribution to features that are absent from the model, while consistency ensures that, if the marginal effect of a feature increases in a modified model, its assigned attribution does not decrease. These properties make SHAP especially useful when interpretability requires formal guarantees rather than purely heuristic explanations.

In the proposed framework, SHAP is not applied directly to the fuzzy decision process, but to a surrogate model that approximates the decision operator f generated by the FAHP–FTOPSIS pipeline. This allows the closeness coefficient to be decomposed into criterion-level contributions without altering the original fuzzy decision mechanism. The resulting SHAP values quantify how each criterion contributes to the predicted ranking score of each alternative and provide the basis for extracting interpretable linguistic rules. In this way, SHAP serves as a bridge between numerical ranking outcomes and the symbolic explanation layer incorporated into the proposed methodology.

2.6. Structural Properties of the Decision Operator

To complement the formal development presented in the previous subsections, this section summarizes several structural properties that help interpret the behavior of the fuzzy MCDM decision operator. These properties are not presented as new theoretical results, but as relevant characteristics associated with the bounded and normalized nature of the FAHP–FTOPSIS procedure used in this study.

First, the closeness coefficient produced by Fuzzy TOPSIS is bounded within the interval [0,1], as it is defined from the relative distances of each alternative to the fuzzy positive and negative ideal solutions. Therefore, the final decision score can be interpreted as a normalized desirability measure, where values closer to one indicate alternatives located nearer to the fuzzy positive ideal solution.

Second, the decision operator is expected to preserve directional consistency with respect to the orientation of the criteria. Improvements in benefit criteria tend to increase the closeness coefficient, whereas improvements in cost criteria, understood as reductions in undesirable values, tend to have the same favorable effect. This property is relevant for interpretability because it allows the contribution of each criterion to be analyzed according to its decision orientation.

Third, within a bounded local neighborhood of the observed alternatives, small perturbations in the input criterion values are expected to produce bounded changes in the

resulting closeness coefficients. This local stability assumption is relevant for the surrogate-based explanation stage because SHAP values are meaningful only if the underlying decision function behaves coherently in the region being approximated.

Taken together, these structural properties support the use of a surrogate model as a local approximation of the fuzzy decision operator. The surrogate and SHAP-based explanation layer do not replace the original FAHP–FTOPSIS mechanism; rather, they provide an interpretable approximation of its input–output behavior within the analyzed decision space.

2.7. Related Work and Research Gap

Fuzzy multi-criteria decision-making has been widely used to support decision problems involving uncertainty, imprecise judgments, and conflicting criteria. Traditional fuzzy MCDM approaches, particularly Fuzzy AHP and Fuzzy TOPSIS, provide structured mechanisms for deriving criteria weights and ranking alternatives under linguistic uncertainty. Previous reviews have shown that fuzzy MCDM has been extensively applied in engineering, management, supplier selection, sustainability assessment, and industrial decision-making contexts [7,8]. These approaches are valuable because they allow decision-makers to incorporate vague and subjective judgments into formal decision models. However, their outputs are usually limited to weights, scores, and rankings, which may not be sufficient when users require explicit explanations of why an alternative is preferred.

A related stream of research has focused on sensitivity and robustness analysis in MCDM. These studies evaluate how changes in criteria weights, input values, or aggregation procedures affect final rankings [17,18]. Robustness analysis is useful because it helps determine whether a decision recommendation remains stable under plausible uncertainty. Nevertheless, sensitivity-oriented approaches generally explain how rankings change under perturbations rather than why a specific ranking emerges from the interaction among criteria. Therefore, they provide important stability evidence, but they do not necessarily produce compact and human-readable explanations of decision logic.

Recent work has also begun to address explainability in MCDM. Explainable TOPSIS and visual MCDM approaches have improved transparency by showing how weights, distances, and aggregation mechanisms influence ranking outcomes [19]. In addition, recent explainable MCDM formulations have proposed new perspectives for making multicriteria decisions more interpretable [12]. These contributions are important because they move MCDM beyond purely numerical ranking outputs. However, many of these approaches still rely mainly on visual inspection, analytical decomposition, or numerical explanation indicators. As a result, the explanation may remain difficult for non-technical decision-makers to operationalize in the form of simple decision rules.

Another relevant research direction combines MCDM with machine learning. Hybrid MCDM–machine learning models have been applied to decision-support problems such as supplier selection, classification, prediction, and risk assessment [20]. These approaches can improve predictive or classification performance and may incorporate explainability tools to interpret model behavior. However, in many cases, the machine learning component may replace, dominate, or obscure the original MCDM logic. This creates a potential interpretability problem because the final decision may become dependent on a predictive model rather than on the formal multicriteria decision structure defined by experts.

SHAP-based explainability methods provide a theoretically grounded way to decompose model outputs into feature-level contributions [11]. SHAP has been widely used in explainable artificial intelligence because it supports both local and global explanation. It can identify how each input variable contributes to a specific prediction and how variables influence model behavior overall. However, SHAP outputs are usually numerical

attribution values or plots. Although these outputs are useful for analysts, they may be less accessible to managers or domain experts who need concise, linguistic, and actionable explanations. Recent SHAP-rule approaches have begun to address this issue by transforming SHAP values into interpretable rule-based explanations [13], but their integration with formal fuzzy MCDM ranking mechanisms remains limited.

Linguistic fuzzy rule-based systems represent another important interpretability tradition. These systems express decision logic through IF–THEN rules, making them attractive for transparent reasoning and communication [21]. Their main advantage is that they provide symbolic explanations that are easy to understand when the rule base remains compact. However, rule generation is often developed independently from formal MCDM ranking mechanisms. Consequently, the resulting rules may not be directly linked to the structure of a fuzzy multicriteria decision operator such as FAHP–FTOPSIS.

Based on this literature, three main gaps can be identified. First, traditional fuzzy MCDM methods handle uncertainty effectively but provide limited explanation of the internal ranking logic. Second, existing explainable MCDM and hybrid MCDM–machine learning approaches often produce visual or numerical explanations, but they do not always transform these explanations into compact linguistic rules. Third, linguistic rule-based explanations are interpretable, but they are not always systematically connected to a formal fuzzy MCDM decision operator. To address these gaps, the present study proposes a unified framework that preserves the FAHP–FTOPSIS ranking structure while adding a surrogate-based SHAP explanation layer and a linguistic IF–THEN rule extraction mechanism. This allows the framework to connect fuzzy ranking, local approximation, numerical attribution, and symbolic explanation within a single decision-support pipeline.

3. Materials and Methods

3.1. Overview of the Explainable Fuzzy MCDM Framework

This study proposes an explainable fuzzy multi-criteria decision-making framework that combines fuzzy multicriteria ranking with a post hoc explainability layer. The aim is to generate not only a ranking of alternatives under uncertainty, but also interpretable explanations that clarify how the evaluation criteria contribute to the final decision score.

Let $A = \{A_1, \ldots, A_m\}$ denote the set of alternatives and $C = \{C_1, \ldots, C_n\}$ the set of evaluation criteria. Expert judgments are collected in linguistic form and transformed into triangular fuzzy numbers, yielding the fuzzy decision matrix $\widetilde{X} = [\widetilde{x}_{ij}]_{m \times n}$, where $\widetilde{x}_{ij}$ represents the fuzzy evaluation of alternative A_i with respect to criterion C_j.

The proposed framework is organized into four sequential stages. First, expert linguistic assessments are converted into fuzzy evaluations and aggregated into a fuzzy decision matrix. Second, criteria weights are obtained through Fuzzy AHP, including a consistency assessment of the defuzzified pairwise comparison matrix. Third, alternatives are ranked using Fuzzy TOPSIS, which yields a closeness coefficient for each alternative. Fourth, a post hoc explainability layer is constructed by training a surrogate model to approximate the FAHP–FTOPSIS decision operator, computing SHAP values, and translating the resulting contribution patterns into interpretable linguistic IF–THEN rules.

The surrogate model is used only as an explanatory approximation of the fuzzy decision operator. It does not replace the original FAHP–FTOPSIS procedure; instead, it supports the interpretation of its input–output behavior. To evaluate this approximation, controlled local perturbations are generated around the original decision profiles, producing augmented decision instances for assessing surrogate fidelity, explanation consistency, and robustness within the bounded decision space of the case study.

Taken together, these stages define an integrated decision pipeline that transforms uncertain linguistic assessments into both quantitative rankings and human-readable

explanations. This structure enables the framework to preserve the uncertainty-handling capacity of fuzzy MCDM while adding an explicit layer of interpretable and auditable decision justification.

3.2. Architecture of the Decision Framework

The overall architecture of the proposed framework is shown in Figure 1. The process starts with the formulation of the decision problem and the collection of expert linguistic evaluations, which are transformed into triangular fuzzy numbers to construct the fuzzy decision matrix. Criteria weights are then derived using Fuzzy AHP, including a consistency assessment of the defuzzified pairwise comparison matrix. The alternatives are subsequently ranked through Fuzzy TOPSIS according to their relative closeness to the fuzzy positive and negative ideal solutions.

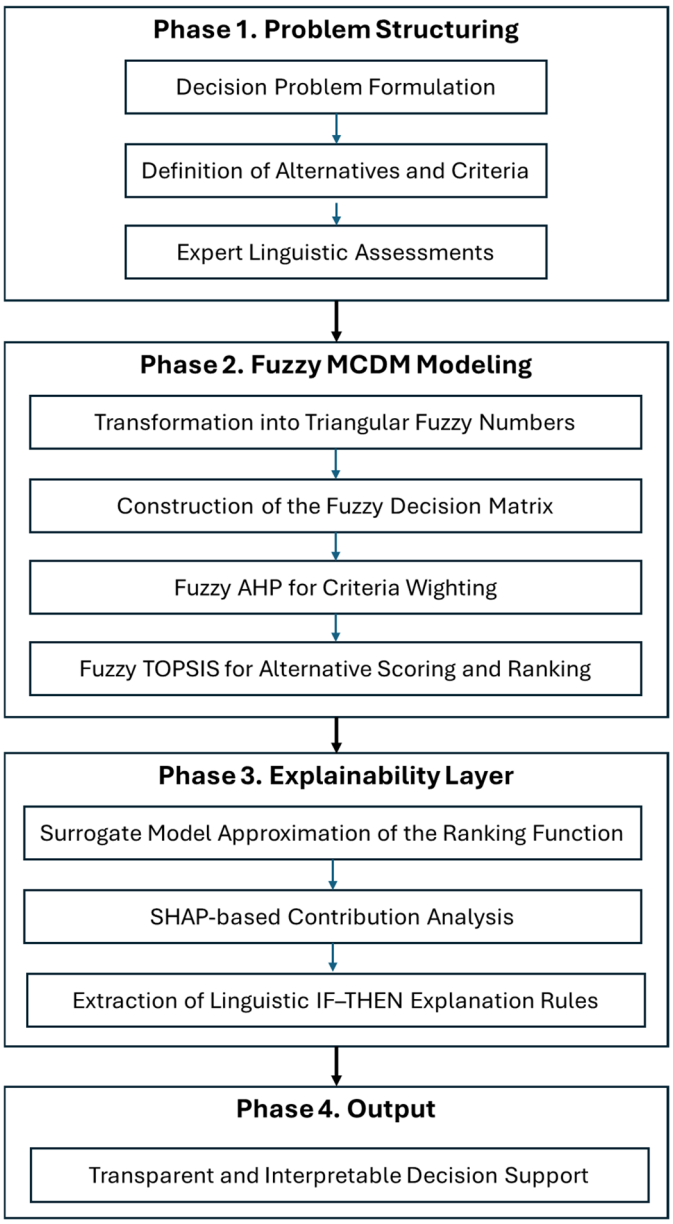

Figure 1. Architecture of the proposed explainable fuzzy multi-criteria decision-making framework, integrating fuzzy evaluation, FAHP weighting, FTOPSIS ranking, surrogate-based approximation, SHAP decomposition, and linguistic rule extraction.

To enhance interpretability, a surrogate model is trained to approximate the input–output behavior of the FAHP–FTOPSIS decision operator within the analyzed decision space. SHAP values are then computed from the surrogate model to quantify the contribution of each criterion to the predicted closeness coefficient. Finally, the resulting contribution patterns are discretized and translated into linguistic IF–THEN rules, providing human-readable explanations of the ranking results.

3.3. Fuzzy Representation and Aggregation of Expert Evaluations

Expert evaluations were collected using linguistic terms and transformed into triangular fuzzy numbers (TFNs) to represent uncertainty in a computationally tractable form. For each alternative A_i, criterion C_j, and expert E_k, the corresponding fuzzy evaluation is denoted by:

$$\tilde{x}_{ij}^{(k)} = \left(l_{ij}^{(k)}, m_{ij}^{(k)}, u_{ij}^{(k)}\right)$$

where $l_{ij}^{(k)}$, $m_{ij}^{(k)}$, and $u_{ij}^{(k)}$ represent the lower, modal, and upper bounds of the assessment provided by expert E_k, respectively.

Since the evaluation process involved three experts, individual fuzzy judgments were aggregated into a collective fuzzy assessment before constructing the final decision matrix. For each alternative A_i and criterion C_j, the aggregated triangular fuzzy number was computed as follows:

$$\tilde{x}_{ij} = (l_{ij}, m_{ij}, u_{ij})$$

where

$$l_{ij} = \min_k\{l_{ij}^{(k)}\}, m_{ij} = \frac{1}{K}\sum_{k=1}^{K} m_{ij}^{(k)}, u_{ij} = \max_k\{u_{ij}^{(k)}\}$$

and $K = 3$ denotes the number of experts. This aggregation rule preserves the uncertainty range expressed by the expert panel by retaining the most conservative lower bound and the most optimistic upper bound, while using the arithmetic mean of the modal values to represent the central tendency of the collective judgment.

All aggregated fuzzy evaluations were then organized into the fuzzy decision matrix:

$$\tilde{X} = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1n} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{x}_{m1} & \tilde{x}_{m2} & \cdots & \tilde{x}_{mn} \end{bmatrix}$$

where each entry $\tilde{x}_{ij}$ represents the aggregated fuzzy performance of alternative A_i under criterion C_j. This matrix serves as the input structure for the subsequent weighting and ranking stages of the proposed framework.

3.4. Criteria Weighting and Consistency Assessment Using Fuzzy AHP

Criteria weights were obtained through fuzzy pairwise comparisons provided by experts using linguistic judgments. These comparisons were represented as triangular fuzzy numbers and organized into the fuzzy pairwise comparison matrix:

$$\tilde{A} = [\tilde{a}_{ij}]_{n\times n}$$

where $\tilde{a}_{ij} = (l_{ij}, m_{ij}, u_{ij})$ denotes the relative importance of criterion C_i with respect to criterion C_j. Reciprocal consistency was enforced as follows:

$$\tilde{a}_{ji} = \left(\frac{1}{u_{ij}}, \frac{1}{m_{ij}}, \frac{1}{l_{ij}}\right), i, j = 1, \ldots, n$$

The fuzzy geometric mean method was used to derive the criteria weights. For each criterion C_i, the fuzzy geometric mean was computed as follows:

$$\tilde{g}_i = \left(\prod_{j=1}^{n} \tilde{a}_{ij}\right)^{1/n}, i = 1, \ldots, n$$

The corresponding normalized fuzzy weight was obtained as follows:

$$\tilde{w}_i = \tilde{g}_i \otimes \left(\bigoplus_{k=1}^{n} \tilde{g}_k\right)^{-1}$$

where $\oplus$ and $\otimes$ denote fuzzy addition and multiplication, respectively. For the subsequent ranking stage, fuzzy weights were defuzzified using the centroid method:

$$w_i = \frac{l_i + m_i + u_i}{3}, i = 1, \ldots, n$$

and then normalized so that:

$$\sum_{i=1}^{n} w_i = 1$$

The resulting crisp weights were incorporated into the weighted fuzzy decision matrix used in the Fuzzy TOPSIS procedure.

To verify the logical coherence of the pairwise comparisons, a consistency analysis was performed after defuzzifying the fuzzy comparison matrix. This step was included because the classical consistency ratio is defined for crisp AHP matrices. First, the maximum eigenvalue $\lambda_{\max}$ was obtained from the defuzzified pairwise comparison matrix. Then, the consistency index was computed as follows:

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

where n denotes the number of criteria. Finally, the consistency ratio was calculated as follows:

$$CR = \frac{CI}{RI}$$

where RI is the random index corresponding to a matrix of order n. A consistency ratio below 0.10 was considered acceptable, indicating that the expert judgments were sufficiently coherent for deriving the criteria weights.

3.5. Ranking of Alternatives Using Fuzzy TOPSIS

Once the criteria weights had been determined, the alternatives were ranked using the Fuzzy TOPSIS method. This method evaluates each alternative according to its relative closeness to the fuzzy positive ideal solution (FPIS) and its distance from the fuzzy negative ideal solution (FNIS).

The fuzzy decision matrix was first normalized to ensure comparability across criteria. Let

$$\tilde{R} = [\tilde{r}_{ij}]_{m \times n}$$

denote the normalized fuzzy decision matrix, where $\tilde{r}_{ij}$ represents the normalized fuzzy performance of alternative A_i with respect to criterion C_j. For benefit criteria, higher values indicate better performance, whereas for cost criteria, lower values are preferred. Accordingly, normalization was performed using standard linear normalization procedures adapted to the orientation of each criterion.

The weighted normalized fuzzy decision matrix was then computed as follows:

$$\tilde{v}_{ij} = \tilde{r}_{ij} \otimes w_j, i = 1, \ldots, m, j = 1, \ldots, n$$

where w_j is the defuzzified and normalized weight associated with criterion C_j, and $\otimes$ denotes scalar multiplication of a triangular fuzzy number.

The fuzzy positive ideal solution and fuzzy negative ideal solution were defined as follows:

$$A^* = \left(\tilde{v}_1^*, \ldots, \tilde{v}_n^*\right), A^- = \left(\tilde{v}_1^-, \ldots, \tilde{v}_n^-\right)$$

where, for each criterion C_j, the ideal values depend on its orientation. For benefit criteria, $\tilde{v}_j^*$ corresponds to the maximum fuzzy value and $\tilde{v}_j^-$ to the minimum fuzzy value. For cost criteria, this definition is reversed.

The distance between two triangular fuzzy numbers $\tilde{a} = (l_a, m_a, u_a)$ and $\tilde{b} = (l_b, m_b, u_b)$ was computed using the vertex method:

$$d(\tilde{a}, \tilde{b}) = \sqrt{\frac{(l_a - l_b)^2 + (m_a - m_b)^2 + (u_a - u_b)^2}{3}}$$

For each alternative A_i, the distances to the fuzzy positive and negative ideal solutions were calculated as follows:

$$D_i^* = \sum_{j=1}^{n} d(\tilde{v}_{ij}, \tilde{v}_j^*), D_i^- = \sum_{j=1}^{n} d(\tilde{v}_{ij}, \tilde{v}_j^-)$$

The final ranking score was obtained through the closeness coefficient:

$$CC_i = \frac{D_i^-}{D_i^* + D_i^-}, i = 1, \ldots, m$$

Alternatives were ranked in descending order of CC_i. Therefore, higher values of CC_i indicate alternatives located closer to the fuzzy positive ideal solution and farther from the fuzzy negative ideal solution. The resulting closeness coefficients were subsequently used as the target values for the surrogate explainability model described in the following subsection.

3.6. Surrogate Explainability Model and Local Perturbation Procedure

Although the Fuzzy TOPSIS procedure provides a ranking of alternatives through the closeness coefficients CC_i, it does not directly reveal the marginal contribution of each criterion to the final score. To address this limitation, a surrogate model was introduced to approximate the decision function induced by the FAHP–FTOPSIS pipeline.

Let

$$x_i = (x_{i1}, \ldots, x_{in})$$

denote the input representation associated with alternative A_i, where each component corresponds to the defuzzified value of criterion C_j used as a predictor in the explainability stage. The surrogate model, denoted by $\hat{g}(x)$, was trained to approximate the mapping:

$$CC_i \approx \hat{g}(x_i), i = 1, \ldots, m$$

The input variables were the defuzzified criterion values of each decision instance, corresponding to cost, quality, delivery reliability, sustainability, risk, and technological capability. The target variable was the closeness coefficient obtained from the FAHP–FTOPSIS pipeline.

A comparative surrogate-model analysis was conducted to identify the model that provided the most appropriate balance between local fidelity and interpretability. Three candidate surrogate models were evaluated: linear regression, a shallow decision tree, and a Random Forest Regressor. Linear regression was included as a transparent additive benchmark, the shallow decision tree as a simple nonlinear and rule-like alternative, and the Random Forest Regressor as a more flexible nonlinear surrogate capable of capturing interaction effects among criteria. All models were trained and evaluated using the same predictor variables, target closeness coefficients, and leave-one-supplier-out cross-validation procedure. The comparison was based on mean absolute error, coefficient of determination, and Spearman rank correlation, allowing both numerical approximation fidelity and ranking preservation to be assessed.

The final surrogate model was selected based on its comparative fidelity and its suitability for SHAP-based explanation. In the final configuration, the Random Forest Regressor was retained as the main surrogate model because it provided the strongest local fidelity while remaining compatible with SHAP-based tree explanations. The model was configured with 500 trees, squared-error criterion, bootstrap sampling enabled, unrestricted maximum depth, a minimum of two samples required to split an internal node, a minimum of one sample per leaf, and a fixed random seed of 42. The selected surrogate was used only as an explanatory approximation of the FAHP–FTOPSIS decision operator and not as a replacement for the original fuzzy multicriteria decision-making procedure.

To avoid fitting the surrogate model directly on the six original supplier alternatives, an augmented local dataset was generated through controlled perturbations of the original decision profiles. This step was introduced because the original alternatives define the empirical decision scenario but do not constitute a sufficiently large sample for training or evaluating a surrogate model in statistical terms. Accordingly, the surrogate was not trained to infer a population-level predictive relationship from six observations; rather, it was trained to approximate the behavior of the FAHP–FTOPSIS decision operator within a bounded local neighborhood around the observed alternatives.

For each supplier alternative, 100 perturbed decision instances were generated by varying the defuzzified criterion values within a $\pm 10\%$ interval around the original values. Perturbations were constrained to remain within the admissible range of each criterion and to preserve the benefit or cost orientation used in the FTOPSIS procedure. Each perturbed instance was then re-evaluated through the complete FAHP–FTOPSIS pipeline to obtain its corresponding closeness coefficient. This process produced 600 decision instances, which were used to assess local surrogate fidelity, SHAP-based explanation consistency, rule extraction, and robustness within the bounded decision space defined by the supplier-selection case study. The $\pm 10\%$ configuration was used as the baseline perturbation setting for surrogate training, SHAP computation, and rule extraction. Additional perturbation magnitudes were evaluated separately in the robustness analysis to examine the sensitivity of ranking outcomes to different levels of local input variation.

To justify the perturbation size used in the local approximation stage, a convergence analysis was conducted using four perturbation configurations: 25, 50, 100, and 200 perturbed instances per original alternative. For each configuration, perturbed profiles were generated within the same ±10% interval, constrained to the admissible range of each criterion and preserving the benefit or cost orientation used in FTOPSIS. Each perturbed profile was then re-evaluated through the complete FAHP–FTOPSIS pipeline to obtain its corresponding closeness coefficient.

The resulting datasets were used to train and evaluate the surrogate model under the same leave-one-supplier-out cross-validation procedure. Convergence was assessed by comparing surrogate fidelity metrics, ranking preservation, and rule stability across perturbation sizes. This analysis was included to determine whether 100 perturbations per alternative provided a sufficiently stable local approximation of the FAHP–FTOPSIS decision operator without unnecessary computational expansion.

Because the perturbed instances were generated around six original supplier profiles, they cannot be treated as statistically independent observations. To avoid information leakage between training and testing subsets, the random 80/20 split was replaced by a leave-one-supplier-out cross-validation strategy. In each fold, all perturbed instances derived from one original supplier were held out as the test set, while the surrogate model was trained on the perturbed instances associated with the remaining suppliers. This grouped validation procedure ensures that no perturbed observations from the same parent supplier appear simultaneously in the training and testing sets.

Surrogate fidelity was evaluated using mean absolute error, coefficient of determination, and Spearman rank correlation across the leave-one-supplier-out folds. These metrics are interpreted strictly as indicators of local approximation fidelity, that is, as evidence of how closely the surrogate reproduces the FAHP–FTOPSIS decision operator within the bounded perturbation neighborhood. They are not interpreted as evidence of population-level predictive generalization, nor are the locally perturbed instances treated as an independent empirical sample.

To complement the local perturbation analysis, a diverse synthetic benchmark was generated across the admissible criterion space. Unlike the ±10% local perturbations, which remain dependent on the six original supplier profiles, the benchmark instances were generated independently by sampling broader combinations of cost, quality, delivery reliability, sustainability, risk, and technological capability while preserving the orientation of benefit and cost criteria. The benchmark consisted of 1000 synthetic profiles generated using Latin hypercube sampling across the admissible range of each criterion, and each profile was evaluated through the complete FAHP–FTOPSIS pipeline to obtain its corresponding closeness coefficient.

The purpose of this benchmark was not to replace validation with real-world data, but to examine whether the proposed explainability framework remains operational under a wider variety of supplier configurations. Therefore, the validation strategy distinguishes between two complementary levels: local leave-one-supplier-out validation based on perturbed empirical profiles, and broader synthetic benchmark evaluation based on independently generated decision profiles.

Once the final surrogate model was selected and trained, SHAP TreeExplainer was used to compute criterion-level contributions to the surrogate-predicted closeness coefficients. Thus, the local surrogate approximation constituted the basis for the SHAP-based decomposition and subsequent linguistic rule extraction stage.

After computing SHAP values from the selected surrogate model, the alignment between the normative weighting structure and the surrogate-based explanation layer was assessed quantitatively. Specifically, Spearman's rank correlation was calculated

between the FAHP-derived criteria weights and the global SHAP importance values. This analysis was conducted using the six criteria included in the case study. Spearman's correlation was selected because the comparison focuses on the ordinal agreement between the criteria importance ranking obtained from FAHP and the ranking derived from global SHAP attributions. A high positive correlation indicates that the criteria considered more important in the FAHP weighting stage are also those exerting greater influence in the surrogate-based explanation layer.

In addition to the perturbation-size convergence analysis, a robustness analysis was conducted using multiple perturbation magnitudes. Specifically, criterion values were perturbed within ±5%, ±10%, and ±15% intervals around each original supplier profile. These magnitudes were selected to represent low, moderate, and relatively stronger local deviations from the baseline evaluations. All perturbations were constrained to remain within the admissible range of each criterion and to preserve the benefit or cost orientation used in the FTOPSIS procedure.

For each perturbation magnitude, the generated profiles were re-evaluated through the complete FAHP–FTOPSIS pipeline and compared with the baseline ranking. Robustness was assessed using two indicators: full-ranking preservation and top-ranked alternative preservation. Full-ranking preservation was defined as the proportion of perturbed scenarios in which the complete ordering of alternatives remained unchanged. Top-ranked alternative preservation was defined as the proportion of scenarios in which the best-ranked alternative remained identical to the baseline result. To account for sampling variability, bootstrapped 95% confidence intervals were computed for both indicators.

3.7. SHAP-Based Rule Extraction

To interpret the surrogate model's predictions, SHAP values were computed for each criterion to quantify its marginal contribution to the predicted ranking score of each decision instance. Formally, the SHAP decomposition expresses the surrogate model output as follows:

$$\hat{g}(x_i) = \phi_0 + \sum_{j=1}^{n} \phi_{ij}$$

where ϕ_{ij} represents the contribution of criterion C_j to the predicted score of decision instance x_i, and ϕ_0 denotes the baseline prediction, corresponding to the expected output of the surrogate model. Positive values of ϕ_{ij} indicate that criterion C_j increases the predicted score, whereas negative values indicate a decreasing effect. This decomposition enables the identification of the most influential criteria in each ranking outcome.

To enhance interpretability, SHAP values were transformed into qualitative contribution levels through a structured rule extraction procedure. First, SHAP values were computed for each decision instance and each criterion using the selected surrogate model. For each criterion, the empirical distribution of absolute SHAP values across the original and perturbed decision instances was used to define percentile-based thresholds. Under the baseline configuration, contributions below the 25th percentile were classified as low, those between the 25th and 75th percentiles as moderate, and those above the 75th percentile as high. The sign of each SHAP value was retained separately to distinguish whether a criterion increased or decreased the predicted closeness coefficient.

Second, the discretized SHAP profiles were combined with the corresponding criterion performance levels. Criteria with high or moderate SHAP magnitudes were considered relevant for the antecedent of a rule, whereas criteria with low SHAP magnitudes were excluded to avoid overly complex and weakly informative rules. This filtering step ensured that the extracted rules focused only on criteria that made a meaningful contribution to the predicted ranking score.

Third, the consequent of each rule was defined according to the predicted closeness coefficient produced by the surrogate model. Predicted scores were grouped into three qualitative priority levels: low, medium, and high. These levels were defined using the empirical distribution of predicted closeness coefficients in the augmented decision dataset so that rule consequents reflected the actual score structure observed in the local decision space.

Finally, recurrent antecedent–consequent patterns were retained as linguistic IF–THEN rules when they satisfied minimum coverage and consistency requirements. Coverage refers to the proportion of decision instances represented by a rule, whereas consistency refers to the proportion of instances satisfying the antecedent that lead to the same consequent. Rules with very low coverage or conflicting consequents were discarded. This procedure reduced the SHAP attribution matrix to a compact set of interpretable rules that summarize stable contribution patterns in the surrogate approximation of the FAHP–FTOPSIS decision operator.

Through this process, numerical SHAP attributions are converted into symbolic decision knowledge. For example, a pattern in which quality and delivery show high positive SHAP contributions, and the predicted closeness coefficient falls in the high-priority group, translates into a rule of the form: "IF quality is high AND delivery is high THEN score is high." In this way, the proposed rule extraction stage provides a practical bridge between local feature-attribution explanations and human-readable decision-support rules.

For each candidate rule r, empirical coverage and consistency were computed to evaluate whether the rule represented a stable pattern in the augmented decision space. Coverage was defined as the proportion of decision instances satisfying the rule antecedent:

$$Coverage(r) = \frac{|\{x_i : x_i \vDash Antecedent(r)\}|}{N}$$

where N denotes the total number of evaluated decision instances. Consistency was defined as the proportion of instances satisfying the antecedent that also matched the same consequent:

$$Consistency(r) = \frac{|\{x_i : x_i \vDash Antecedent(r) \wedge y_i = Consequent(r)\}|}{|\{x_i : x_i \vDash Antecedent(r)\}|}$$

For the final rule set, coverage was computed as the proportion of instances covered by at least one retained rule, whereas consistency was computed as the proportion of covered instances whose consequent matched the predicted priority class. This distinction was introduced because individual-rule metrics and rule-set-level metrics capture different aspects of interpretability: the former evaluates the reliability of each rule, while the latter evaluates the explanatory coverage and coherence of the retained rule set as a whole.

To provide a reference point for evaluating the extracted rule set, a trivial majority-class rule was used as a baseline. This baseline assigns all decision instances to the most frequent qualitative priority class in the augmented decision space without using any criterion-level antecedents. Because this rule has no antecedent, it covers all instances by construction; however, its consistency is limited to the prevalence of the majority class and it provides no criterion-level explanatory structure. Therefore, the baseline serves as a minimal benchmark for assessing whether the SHAP-guided rules offer explanatory value beyond reproducing the dominant class.

In addition, bootstrapped 95% confidence intervals were estimated for the coverage and consistency of the extracted rule set. Bootstrap samples were generated by resampling the augmented decision instances with replacement. For each bootstrap sample, the coverage and consistency of the retained rule set were recalculated. This procedure was

used to evaluate the stability of the rule-quality indicators and to avoid relying exclusively on point estimates.

To assess whether the extracted rules were sensitive to the percentile thresholds used in SHAP discretization, a threshold sensitivity analysis was conducted after the baseline rule set had been generated. The 25th/75th percentile scheme was retained as the baseline configuration because it provides a balanced separation between low, moderate, and high contribution magnitudes. Two alternative schemes were also tested: 20th/80th percentiles, representing a stricter classification of high and low contributions, and 30th/70th percentiles, representing a more permissive classification.

For each threshold configuration, SHAP magnitudes were reclassified, the rule extraction procedure was repeated, and the resulting rule set was evaluated in terms of number of retained rules, empirical coverage, consistency, and overlap with the baseline rule set. Rule overlap was computed as the proportion of baseline rules that were preserved under each alternative threshold scheme. This analysis was included to verify whether the compactness and consistency of the extracted rules reflected stable attribution patterns rather than artifacts of a single discretization choice.

The general procedure for the proposed framework is summarized in Section 3.8.

3.8. Algorithm 1. Explainable Fuzzy Multi-Criteria Decision-Making Framework

The overall procedure of the proposed framework is summarized in Algorithm 1. The algorithm takes as input the set of alternatives, the set of criteria, expert linguistic evaluations, and expert pairwise comparison judgments. It returns the final ranking of alternatives and a set of interpretable linguistic IF–THEN rules explaining the ranking outcomes. Unlike a purely descriptive workflow, the algorithm explicitly separates the fuzzy multicriteria decision-making layer from the surrogate-based explainability layer.

Algorithm 1. Explainable Fuzzy Multi-Criteria Decision-Making Framework:

Input: Set of alternatives $A = \{A_1, A_2, \ldots, A_m\}$; set of criteria $C = \{C_1, C_2, \ldots, C_n\}$; expert linguistic evaluations of alternatives; expert pairwise comparison judgments for criteria; linguistic-to-fuzzy conversion scale; number of local perturbations per alternative N_p; perturbation range δ; minimum rule coverage threshold γ; and minimum rule consistency threshold κ.

Output: Final ranking of alternatives; surrogate fidelity metrics; global and local SHAP explanations; and interpretable linguistic IF–THEN decision rules.

Procedure:

1. Initialize the fuzzy decision matrix $\widetilde{X}$, the criteria weight vector w, the closeness coefficient vector CC, the perturbed dataset D_p, the SHAP matrix Φ, and the rule set R.
2. For each expert evaluation of alternative A_i under criterion C_j:
   - 2.1. Convert the linguistic evaluation into a triangular fuzzy number.
   - 2.2. Store the resulting fuzzy value in the individual expert decision matrix.
3. For each alternative A_i and criterion C_j:
   - 3.1. Aggregate the individual expert fuzzy evaluations.
   - 3.2. Construct the collective fuzzy decision matrix $\widetilde{X}$.
4. Convert expert pairwise comparison judgments into triangular fuzzy numbers and construct the fuzzy pairwise comparison matrix.

Algorithm 1. *Cont.*

5. Apply Fuzzy AHP to obtain the fuzzy criteria weights.
6. Defuzzify and normalize the criteria weights to obtain the crisp weight vector w.
7. Assess the consistency ratio of the defuzzified pairwise comparison matrix.
    - 7.1. If the consistency ratio exceeds the accepted threshold, revise the pairwise comparisons.
    - 7.2. Else, retain the criteria weights for the ranking stage.
8. Normalize the fuzzy decision matrix according to the orientation of each criterion.
9. Construct the weighted normalized fuzzy decision matrix using the criteria weights w.
10. Determine the fuzzy positive ideal solution and fuzzy negative ideal solution.
11. For each alternative A_i:
    - 11.1 Compute its distance to the fuzzy positive ideal solution.
    - 11.2 Compute its distance to the fuzzy negative ideal solution.
    - 11.3 Calculate its closeness coefficient CC_i.
12. Rank the original alternatives in descending order of their closeness coefficients.
13. For each original alternative A_i:
    - 13.1 Generate N_p locally perturbed decision profiles within the interval $\pm\delta$.
    - 13.2 Preserve the admissible range and benefit/cost orientation of each criterion.
    - 13.3 Re-evaluate each perturbed profile through the complete FAHP–FTOPSIS pipeline.
    - 13.4 Store the perturbed criterion values and corresponding closeness coefficients in D_p.
14. Generate an additional diverse synthetic benchmark across the admissible criterion space and evaluate each synthetic profile through the FAHP–FTOPSIS pipeline.
15. Define the surrogate-model dataset using criterion values as predictors and FAHP–FTOPSIS closeness coefficients as target values.
16. Train and compare candidate surrogate models, including linear regression, a shallow decision tree, and a Random Forest Regressor.
17. Evaluate surrogate fidelity using leave-one-supplier-out cross-validation based on MAE, R^2, and Spearman rank correlation.
18. Select the surrogate model that provides the most appropriate balance between local fidelity and interpretability.
19. Compute SHAP values for the selected surrogate model to obtain criterion-level contributions for each decision instance.
20. Calculate global SHAP importance values by aggregating absolute SHAP values across decision instances.
21. Assess the alignment between FAHP weights and global SHAP importance values using Spearman's rank correlation.
22. For each threshold scheme considered for SHAP discretization:
    - 22.1. Define percentile-based thresholds for low, moderate, and high SHAP contribution magnitudes.
    - 22.2. Retain the sign of each SHAP value to distinguish positive and negative effects.
    - 22.3. Transform numerical SHAP values into linguistic contribution categories.

Algorithm 1. *Cont.*

23. For each decision instance:
    - 23.1. Select criteria with moderate or high SHAP contribution magnitudes as candidate rule antecedents.
    - 23.2. Assign the predicted closeness coefficient to a qualitative priority level.
    - 22.3. Form a candidate IF–THEN rule linking contribution patterns to the priority level.
24. For each candidate rule r:
    - 24.1. Compute its empirical coverage.
    - 24.2. Compute its empirical consistency.
    - 24.3. If coverage $\geq \gamma$ and consistency $\geq \kappa$, retain r in the final rule set R.
    - 24.4. Else, discard r.
25. Compare the retained rule sets across alternative SHAP discretization thresholds to assess rule stability.
26. Return the final alternative ranking, surrogate fidelity metrics, SHAP explanations, threshold sensitivity results, and final IF–THEN rule set R.

This algorithm formalizes the complete operational sequence of the proposed framework. Steps 1–12 correspond to the fuzzy multicriteria decision-making layer, where expert linguistic evaluations are transformed into fuzzy rankings. Steps 13–18 define the surrogate approximation layer, including local perturbation, synthetic benchmarking, model comparison, and grouped validation. Steps 19–26 correspond to the explainability layer, where SHAP values are computed, discretized, and transformed into interpretable linguistic rules. This structure improves reproducibility by making explicit the computational flow from fuzzy input data to final ranking explanations.

4. Results

4.1. Case Study Instantiation

To operationalize the proposed framework, the empirical validation was conducted on a supplier selection problem under uncertainty, consistent with the methodological design presented in the previous section.

A set of six candidate suppliers was considered:

$$A = \{A_1, A_2, A_3, A_4, A_5, A_6\}$$

The evaluation was performed using six criteria representing the main economic, operational, and strategic dimensions of supplier assessment:

- Cost (C_1)—cost criterion.
- Quality (C_2)—benefit criterion.
- Delivery reliability (C_3)—benefit criterion.
- Sustainability (C_4)—benefit criterion.
- Risk (C_5)—cost criterion.
- Technological capability (C_6)—benefit criterion.

This configuration was selected because it reflects a realistic multicriteria purchasing context in which decision-makers must simultaneously balance efficiency, operational reliability, innovation capacity, and exposure to adverse conditions.

The criteria were selected because they represent the main dimensions commonly involved in supplier evaluation decisions. Cost captures the economic dimension of

the purchasing decision, quality and delivery reliability reflect operational performance, sustainability incorporates environmental and social considerations, risk accounts for potential exposure to uncertainty or disruption, and technological capability represents the supplier's capacity to support innovation and process improvement. Therefore, the selected criteria provide a balanced representation of economic, operational, strategic, and sustainability-related aspects of supplier selection.

A panel of three experts provided linguistic evaluations using a five-level scale. The experts were selected based on their knowledge of procurement processes, operations management, and multicriteria decision analysis, ensuring that the evaluations reflected both practical decision-making experience and methodological understanding of supplier assessment. Their individual fuzzy evaluations were aggregated into a single fuzzy decision matrix using the procedure described in Section 3.3, which combines the minimum lower bound, the average modal value, and the maximum upper bound across experts for each alternative–criterion pair.

After the linguistic-to-fuzzy transformation and aggregation processes, the resulting fuzzy decision matrix $\widetilde{X}$ constituted the input to the integrated FAHP–FTOPSIS decision pipeline. This empirical instantiation provides the basis for assessing not only the ranking performance of the proposed framework, but also the fidelity, interpretability, and robustness of its explanatory layer.

Based on this empirical configuration, the fuzzy MCDM pipeline was applied to derive criteria weights and ranking results, as presented in the following subsection.

Although the case study considers only six original supplier alternatives, this number should be understood as the size of the illustrative decision problem rather than as a statistical sample for predictive inference. The purpose of the case study is to demonstrate the operational sequence of the proposed framework in a realistic supplier-selection context. Therefore, the statistical interpretation of the surrogate model metrics is deliberately restricted: these metrics evaluate local approximation fidelity within the augmented perturbation space, not generalizable predictive performance from the six original alternatives alone. Nevertheless, the framework is structurally scalable, since the fuzzy weighting, ranking, surrogate modeling, and SHAP-based explanation stages can be applied to larger sets of alternatives and criteria. In such applications, however, computational cost, perturbation design, and rule-set complexity must be carefully managed.

4.2. Fuzzy MCDM Results

Before deriving the final criteria weights, the defuzzified pairwise comparison matrix was examined for logical coherence. For the six-criteria matrix, the maximum eigenvalue was $\lambda_{\max} = 6.397$, yielding a consistency index of $CI = 0.079$. Using the corresponding random index $RI = 1.24$, the resulting consistency ratio was $CR = 0.064$. Since this value is below the commonly accepted threshold of 0.10, the pairwise comparison judgments were considered sufficiently consistent for the FAHP weighting procedure.

After applying the Fuzzy AHP procedure and defuzzification, the normalized criteria weights are presented in Table 1.

These weights indicate that quality and cost dominate the decision structure, followed by delivery and technological capability. More specifically, quality (0.22) emerges as the most influential criterion, suggesting that decision-makers prioritize performance-related attributes. Cost (0.18) follows closely, indicating that economic considerations remain central but do not override quality-driven preferences.

The remaining criteria exhibit a relatively balanced distribution, reinforcing the inherently multicriteria nature of the problem, where trade-offs between competing dimensions are required rather than single-criterion optimization.

Table 1. Criteria weights obtained through FAHP.

Criterion	Weight
Cost (C_1)	0.18
Quality (C_2)	0.22
Delivery (C_3)	0.17
Sustainability (C_4)	0.14
Risk (C_5)	0.13
Technology (C_6)	0.16

The Fuzzy TOPSIS closeness coefficients are summarized in Table 2. The resulting ranking is: $A_4 > A_2 > A_6 > A_5 > A_1 > A_3$

Table 2. Closeness coefficients (Fuzzy TOPSIS).

Alternative	CC_i
A_1	0.62
A_2	0.71
A_3	0.55
A_4	0.78
A_5	0.64
A_6	0.69

The results show that A_4 is the most suitable alternative. Importantly, its superiority is not driven by extreme performance in a single criterion, but by a balanced profile across benefit criteria combined with acceptable cost–risk trade-offs.

This pattern confirms that the fuzzy MCDM operator captures a compensatory decision logic, where strengths in key criteria offset moderate weaknesses in others. Additionally, the observed spread in closeness coefficients (0.55–0.78) indicates that the model provides sufficient discrimination among alternatives, avoiding both ranking degeneracy and excessive sensitivity.

4.3. Surrogate Model Performance

To approximate the decision function induced by the fuzzy MCDM pipeline, three candidate surrogate models were evaluated using the augmented local dataset described in Section 3.6. The dataset consisted of 600 perturbed decision instances generated around the six original supplier profiles. Each perturbed instance was processed through the FAHP–FTOPSIS pipeline, and the resulting closeness coefficient was used as the target value for surrogate training and evaluation. Therefore, the surrogate models were not trained directly on the six original alternatives, nor were the reported metrics computed as evidence of statistical generalization from six empirical observations. Instead, the metrics evaluate how faithfully each surrogate approximates the FAHP–FTOPSIS decision operator within the bounded perturbation space.

The evaluated models were linear regression, a shallow decision tree, and a Random Forest Regressor. Linear regression was included as a transparent additive benchmark, while the shallow decision tree represented a simple nonlinear and rule-like alternative. The Random Forest Regressor was evaluated as a more flexible nonlinear surrogate capable of capturing interaction effects among criteria. All models were trained and evaluated using the same predictor variables, target closeness coefficients, and leave-one-supplier-out cross-validation procedure.

The comparative performance metrics are reported in Table 3. These metrics should be interpreted as indicators of local surrogate fidelity, that is, as measures of how closely each

surrogate approximates the behavior of the original fuzzy decision operator within the bounded perturbation space. They should not be interpreted as results from an independent synthetic benchmark or as evidence of population-level predictive generalization.

Table 3. Comparative surrogate model performance under leave-one-supplier-out cross-validation.

Surrogate Model	MAE	R^2	Spearman ρ
Linear regression	0.048	0.757	0.771
Shallow decision tree	0.031	0.890	0.886
Random Forest Regressor	0.018	0.961	0.943

The comparison shows that the Random Forest Regressor achieved the highest local fidelity among the evaluated surrogate models. Linear regression provided a useful transparent baseline, but its lower performance indicates that the FAHP–FTOPSIS decision operator was not fully captured by a purely additive linear approximation. The shallow decision tree offered a more interpretable structure than the Random Forest, but its reduced fidelity and lower ranking-preservation capacity limited its suitability as the main explanation model.

Based on this comparison, the Random Forest Regressor was retained as the principal surrogate model because it provided the best balance between local approximation fidelity and compatibility with SHAP-based explanation. Importantly, the Random Forest was not used as a decision model replacing FAHP–FTOPSIS; it was used only as an explanatory proxy for decomposing the behavior of the fuzzy decision operator within the analyzed decision space.

The results indicate that the selected surrogate closely approximates the FAHP–FTOPSIS decision operator in the locally generated decision space. The low MAE reflects small deviations between the surrogate-predicted and FTOPSIS-derived closeness coefficients, while the high Spearman correlation indicates that the ordinal structure of the ranking is largely preserved within the perturbation-based dataset. This distinction is central to the proposed framework: because SHAP values are computed from the surrogate model, the usefulness of the explanation layer depends on the extent to which the surrogate reproduces the behavior of the fuzzy decision operator in the analyzed region of the input space. Within this local fidelity perspective, the surrogate model acts as an explanatory approximation rather than as an independent predictive model.

4.4. SHAP-Based Explanation

The global SHAP importance values are presented in Table 4, together with the FAHP-derived criteria weights. This comparison was included to evaluate whether the surrogate-based explanation layer preserved the relative importance structure embedded in the original fuzzy MCDM model.

Table 4. FAHP weights and global SHAP importance of criteria.

Criterion	FAHP Weight	Global SHAP Importance	FAHP Rank	SHAP Rank
Quality (C_2)	0.22	0.24	1	1
Cost (C_1)	0.18	0.21	2	2
Delivery (C_3)	0.17	0.18	3	3
Technology (C_6)	0.16	0.15	4	4
Sustainability (C_4)	0.14	0.12	5	5
Risk (C_5)	0.13	0.10	6	6

Note: Spearman's rank correlation between FAHP weights and global SHAP importance values was $\rho = 1.00$ ($p < 0.01$). Given the small number of criteria, the p-value should be interpreted only as a descriptive inferential indicator.

To complement the aggregated values reported in Table 4, Figure 2 presents the SHAP summary plot of the selected surrogate model. This visualization shows both the relative importance of the criteria and the direction of their contribution to the predicted closeness coefficient.

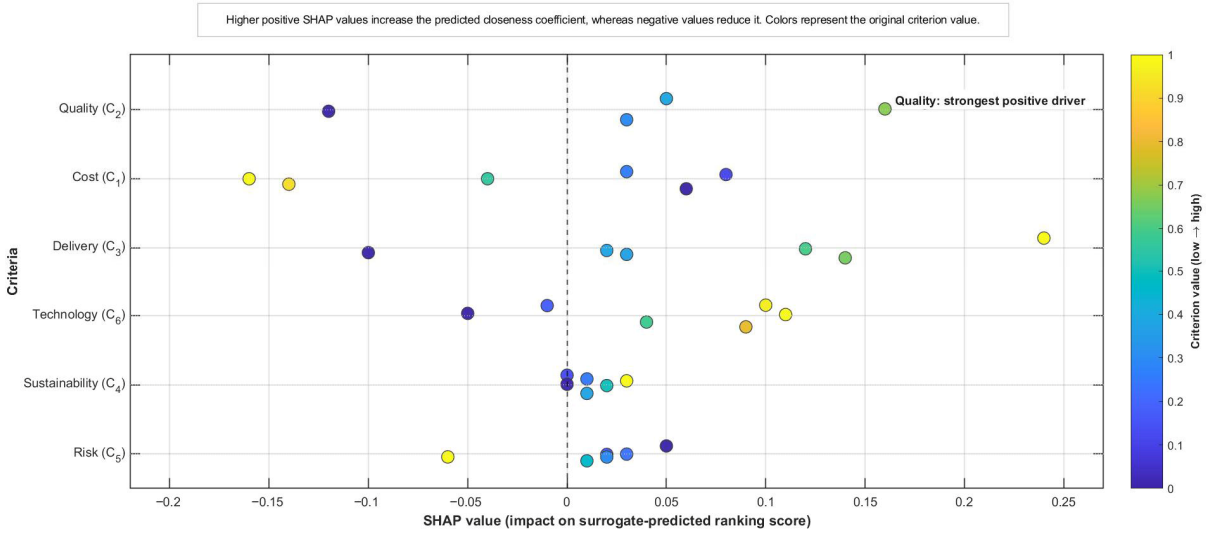

Figure 2. SHAP summary plot of criterion contributions to the surrogate-predicted ranking score.

As shown in Table 4 and Figure 2, quality, cost, and delivery are the most influential criteria in the surrogate model. More importantly, the full ordinal ranking of criteria is identical in the FAHP weights and the global SHAP importance values. The Spearman correlation confirms a perfect positive ordinal association between both importance structures, with $\rho = 1.00$ ($p < 0.01$). This indicates that the criteria considered most relevant in the normative weighting stage were also those exerting the strongest influence in the surrogate-based explanation layer.

This result supports the internal coherence of the proposed framework because the SHAP-based explanation layer does not introduce an alternative criterion-priority structure, but reproduces the relative importance logic embedded in the original FAHP–FTOPSIS model. However, the result should be interpreted primarily as evidence of ordinal alignment within the present case study, not as a general statistical claim. Since the correlation is based on six criteria, this alignment should be reassessed in larger applications involving more criteria, alternatives, and heterogeneous decision contexts.

At the local level, the SHAP decomposition also clarifies why specific alternatives obtain their ranking positions. For the top-ranked alternative, A_4, the main positive contributions come from quality and delivery, while technological capability provides a moderate positive contribution. Cost has a negative effect, whereas sustainability contributes only marginally. This pattern shows that the superiority of A_4 is not due to uniformly high performance across all criteria, but to a favorable configuration in which strong benefit-criterion performance offsets the adverse effect of cost.

From an interpretability perspective, this decomposition transforms the closeness coefficient from a single numerical score into a structured explanation of the ranking outcome. It shows not only which alternative is preferred, but also which criteria drive that preference and how trade-offs are resolved at the alternative level.

4.4.1. Perturbation-Size Convergence Analysis

A convergence analysis was conducted to justify the number of local perturbations generated around each original supplier profile. Four perturbation sizes were compared:

25, 50, 100, and 200 perturbed instances per alternative. For each configuration, the perturbed profiles were generated within the same ±10% interval and re-evaluated through the complete FAHP–FTOPSIS pipeline. The resulting datasets were then used to assess surrogate fidelity, ranking preservation, and rule stability.

Table 5 reports the convergence results. The results show that surrogate fidelity and ranking stability improved as the number of perturbations increased from 25 to 50 and from 50 to 100. However, the improvement from 100 to 200 perturbations was marginal, indicating that the local approximation had largely stabilized at 100 perturbations per alternative. This supports the use of 100 perturbations as a sufficient configuration for the present case study, balancing approximation stability and computational efficiency.

Table 5. Stability of surrogate fidelity and rule consistency across perturbation sizes.

Perturbations per Alternative	Total Perturbed Instances	MAE	R^2	Spearman ρ	Ranking Preservation	Rule-Set Overlap
25	150	0.034	0.88	0.86	0.82	0.62
50	300	0.024	0.93	0.91	0.87	0.83
100	600	0.018	0.96	0.94	0.89	1.00
200	1200	0.017	0.96	0.95	0.90	0.98

The convergence pattern indicates that the main performance indicators stabilized at 100 perturbations per alternative. Although the 200-perturbation configuration produced a larger augmented dataset, it did not substantially improve surrogate fidelity or rule stability compared with the 100-perturbation configuration. Therefore, the baseline configuration of 100 perturbations per alternative was retained for the subsequent SHAP analysis and rule extraction stages.

4.4.2. Grouped Cross-Validation and Synthetic Benchmark Evaluation

To account for the dependence structure introduced by local perturbations, surrogate fidelity was evaluated using leave-one-supplier-out cross-validation. In each fold, all perturbed instances associated with one supplier were excluded from training and used exclusively for testing. This procedure provides a more conservative assessment than a random train–test split because the model is evaluated on supplier profiles that were not represented in the training fold.

The leave-one-supplier-out results indicate that the surrogate model maintains adequate local fidelity when evaluated under grouped validation. This supports its use as an explanatory approximation of the FAHP–FTOPSIS operator, while avoiding the overinterpretation of perturbed instances as independent empirical observations.

In addition, the framework was evaluated using a diverse synthetic benchmark generated across the admissible criterion space. The benchmark analysis confirmed that the proposed pipeline can be applied beyond the immediate neighborhood of the six illustrative suppliers. However, the results also show that explanation stability depends on the structure of the sampled decision space, reinforcing the need to distinguish local robustness from broader generalizability.

After comparing the candidate surrogate models, the Random Forest Regressor was selected for the SHAP-based explanation stage. Its fidelity was then examined under the grouped validation setting and the broader synthetic benchmark. Table 6 summarizes these results.

4.5. Rule Extraction

The rules reported in Table 7 were obtained by applying the SHAP-guided rule extraction procedure described in Section 3.7. First, SHAP values were discretized into

low, moderate, and high contribution levels using criterion-specific percentile thresholds. Second, only criteria with moderate or high contribution magnitudes were retained in the antecedents, while low-contribution criteria were excluded to preserve rule compactness. Third, the predicted closeness coefficients were grouped into low-, medium-, and high-priority levels to define the rule consequents. Finally, recurrent antecedent–consequent patterns were retained when they exhibited adequate empirical coverage and consistency within the augmented decision instances. This procedure ensured that the extracted rules represented stable explanation patterns rather than isolated individual cases.

Table 6. Fidelity of the selected Random Forest surrogate under grouped validation and synthetic benchmark evaluation.

Validation Setting	MAE	R^2	Spearman ρ
Leave-one-supplier-out CV	0.018	0.96	0.94
Synthetic benchmark	0.035	0.91	0.96

Table 7. Extracted linguistic decision rules.

Rule	Description
R1	IF quality is high AND delivery is high THEN score is high
R2	IF cost is low AND risk is low THEN score is high
R3	IF quality is medium AND delivery is low THEN score is medium
R4	IF cost is high AND risk is high THEN score is low

The extracted rules reveal two dominant decision regimes. The first is a performance-driven regime, where high quality and delivery jointly lead to superior outcomes. The second is a risk–cost control regime, where low cost and low risk reinforce favorable evaluations. Conversely, the combination of high cost and high risk defines an unfavorable region of the decision space associated with low scores.

Under the baseline 25th/75th-percentile discretization scheme, the final SHAP-guided rule set achieved 83% coverage and 0.91 consistency. To contextualize these values, the rule set was compared against a trivial majority-class rule that assigns all instances to the most frequent priority level without using criterion-level antecedents. The majority-class baseline provides a minimal reference because it captures class prevalence but does not offer explanatory conditions or decision logic.

Bootstrap resampling was also used to estimate 95% confidence intervals for the rule-quality indicators. The SHAP-guided rule set maintained high and stable performance, with coverage and consistency values remaining within narrow confidence intervals. Compared with the majority-class baseline, the extracted rules provided a more informative explanation structure by linking priority outcomes to explicit criterion conditions, rather than merely reproducing the dominant class. Therefore, the reported coverage and consistency values support the interpretability of the extracted rules as structured explanations rather than trivial class summaries.

By translating numerical attribution patterns into interpretable IF–THEN statements, the rule layer supports managerial decision-making with transparent, actionable selection criteria (Table 8).

As shown in Table 8, the SHAP-guided rule set provides a stronger explanation structure than the majority-class baseline. Although the baseline reflects the most frequent priority level, it does not identify the criterion configurations associated with each decision outcome. In contrast, the SHAP-guided rules maintain high coverage and consistency while preserving explicit antecedent–consequent relationships. The bootstrapped confi-

dence intervals further indicate that the reported rule-quality indicators are stable under resampling.

Table 8. Rule-quality comparison with majority-class baseline and bootstrap confidence intervals.

Rule Model	Coverage	95% CI for Coverage	Consistency	95% CI for Consistency
Majority-class baseline	1.00	1.00–1.00	0.43	0.39–0.47
SHAP-guided rule set	0.83	0.80–0.86	0.91	0.88–0.93

4.6. Sensitivity Analysis of SHAP Discretization Thresholds

To verify whether the extracted linguistic rules were stable across the percentile thresholds used to discretize SHAP values, three threshold configurations were compared: 20th/80th, 25th/75th, and 30th/70th percentiles. The 25th/75th scheme served as the baseline configuration, while the 20th/80th and 30th/70th schemes were used to examine stricter and more permissive thresholds for SHAP contribution magnitudes, respectively.

Table 9 summarizes the sensitivity analysis. Across the three configurations, the main rule structure remained stable. In particular, the dominant rules involving quality and delivery reliability as positive drivers of high scores, and cost and risk as limiting factors, were preserved under all threshold schemes. The stricter 20th/80th configuration produced a slightly more selective rule set, whereas the more permissive 30th/70th configuration retained additional marginal antecedents. However, these variations did not alter the central explanatory logic of the extracted rules.

Table 9. Rule-set robustness across SHAP discretization schemes.

Threshold Scheme	Number of Retained Rules	Coverage	Consistency	Overlap with Baseline Rule Set
20th/80th percentiles	3	0.76	0.93	0.75
25th/75th percentiles	4	0.83	0.91	1.00
30th/70th percentiles	5	0.87	0.89	1.00

These results indicate that the rule set is not dependent on a single arbitrary threshold choice. Although coverage and consistency vary slightly across threshold schemes, the core rules remain stable, supporting the interpretation that the extracted linguistic rules summarize recurrent SHAP contribution patterns rather than threshold-specific artifacts.

The baseline 25th/75th configuration achieved 83% coverage and 0.91 consistency. The alternative threshold schemes produced comparable rule structures, with only minor changes in coverage and consistency. The overlap values indicate that the principal rules were preserved when the discretization thresholds were modified. Therefore, the rule extraction procedure can be considered stable within the tested percentile range.

4.7. Robustness Analysis

To evaluate the stability of the proposed framework under local input uncertainty, robustness was assessed across three perturbation magnitudes: $\pm 5\%$, $\pm 10\%$, and $\pm 15\%$. These magnitudes represent low, moderate, and relatively stronger deviations around the original supplier profiles. For each magnitude, perturbed profiles were generated while preserving the admissible range and benefit/cost orientation of each criterion, and each profile was re-evaluated through the complete FAHP–FTOPSIS pipeline.

Two robustness indicators were calculated. Full-ranking preservation measures the proportion of perturbation scenarios in which the complete ordering of alternatives remained unchanged relative to the baseline ranking. Top-ranked alternative preservation

measures the proportion of scenarios in which the best-ranked supplier remained the same as in the original ranking. Bootstrapped 95% confidence intervals were estimated for both indicators to account for sampling variability in the perturbation-based robustness assessment.

Table 10 reports the robustness results across perturbation magnitudes. The results show that ranking stability decreases gradually as perturbation magnitude increases. Under the ±5% condition, both the complete ranking and the top-ranked alternative remain highly stable, indicating that minor input variations have little effect on the decision outcome. Under the ±10% condition, the full ranking is preserved in 89% of perturbation scenarios, while the top-ranked alternative remains unchanged in 94% of cases. Under the ±15% condition, full-ranking preservation decreases further, but the top-ranked alternative remains comparatively stable. This pattern indicates that the framework is locally robust, particularly in identifying the best supplier.

Table 10. Robustness analysis across perturbation magnitudes.

Perturbation Magnitude	Full-Ranking Preservation	95% CI	Top-Ranked Alternative Preservation	95% CI
±5%	0.96	0.94–0.98	0.99	0.98–1.00
±10%	0.89	0.87–0.91	0.94	0.92–0.96
±15%	0.81	0.78–0.84	0.88	0.85–0.91

These findings are consistent with the structural properties discussed in Section 2.6, particularly the expectation that bounded input perturbations should yield bounded variations in closeness coefficients. At the same time, the decrease in full-ranking preservation under larger perturbations shows that robustness is magnitude-dependent. Therefore, the robustness results should be interpreted as evidence of stability within the tested perturbation range, rather than as proof of invariance under arbitrary input changes.

This stability also reinforces the reliability of the explanation layer. Since SHAP explanations depend on local variations in the input space, a stable decision function increases confidence that the derived explanations reflect meaningful criterion-level relationships rather than artifacts of random noise. From an XAI perspective, this is relevant because explanation quality depends not only on interpretability, but also on the stability of the decision behavior being explained.

5. Discussion

The results support the central premise of the proposed framework: fuzzy multicriteria decision-making can be integrated with explainability techniques to produce not only rankings under uncertainty, but also transparent and interpretable decision-support mechanisms. This finding is consistent with the broader evolution of fuzzy MCDM research, which has moved from purely ranking-oriented formulations toward models that preserve uncertainty handling while improving interpretability, traceability, and decision-support quality in complex environments [7–9]. In this sense, the proposed framework contributes to applied AI by combining uncertainty modeling, surrogate approximation, and interpretable knowledge extraction within a unified decision architecture.

5.1. Interpretation of Main Findings

From an applied artificial intelligence perspective, the results demonstrate the practical viability of integrating fuzzy reasoning with model-based explanation techniques. The fuzzy MCDM layer generated a stable and discriminative ranking of alternatives, with clear differentiation among candidates based on their closeness coefficients. The dominance

of alternative A_4 was not driven by extreme performance in a single criterion, but by a balanced configuration across benefit criteria combined with acceptable cost–risk trade-offs. This confirms that the proposed approach preserves the compensatory logic of multicriteria decision-making, where trade-offs between criteria are systematically resolved. This interpretation is consistent with the theoretical foundations and widespread applications of TOPSIS and fuzzy TOPSIS, which evaluate alternatives through their relative closeness to ideal solutions while accommodating compensatory behavior among conflicting criteria [6,7,22,23].

However, these empirical results should be interpreted within the scope of the designed decision scenario. The six original alternatives instantiate the supplier-selection problem, whereas the augmented perturbation-based dataset supports the evaluation of local surrogate fidelity and explanation stability. Consequently, the findings demonstrate the methodological feasibility and internal coherence of the framework, but they do not constitute broad statistical validation across supplier-selection populations or industries.

A key finding is the quantitative alignment between the FAHP-derived criteria weights and the SHAP-based global importance values. The Spearman correlation analysis showed a perfect positive ordinal association between both importance structures, indicating that the criteria identified as more relevant in the normative weighting stage were also those exerting greater influence in the surrogate-based explanation layer. This strengthens the internal coherence of the proposed framework because the explanation layer does not introduce an alternative criterion-priority structure, but reproduces the ranking logic embedded in the original FAHP–FTOPSIS model. Nevertheless, this result should be interpreted cautiously, as the correlation is based on six criteria from a single illustrative supplier-selection scenario. Therefore, the framework does not sacrifice coherence for interpretability; rather, it preserves consistency between the decision structure and its explanation. This is methodologically relevant because a common concern in explainable decision systems is that the explanation layer may diverge from the logic of the original model. In contrast, the present results suggest convergence between fuzzy weighting and feature-attribution mechanisms, in line with recent efforts to make MCDM methods explainable without sacrificing structural fidelity [8,11,12].

At the local level, the SHAP decomposition provided an interpretable explanation of individual ranking outcomes. For the top-ranked alternative, positive contributions from quality and delivery offset the negative effect of cost, demonstrating how the model operationalizes trade-offs at the alternative level. This reinforces the view that explainability in multicriteria contexts should not be limited to global importance measures, but should also capture local decision mechanisms. This finding is aligned with the rationale of SHAP-based analysis, according to which local additive attributions explain how individual variables contribute to a specific output, while aggregated attributions support a broader understanding of model behavior [9–11]. From an operational perspective, this capability improves usability because stakeholders can understand not only which alternative is preferred, but also the trade-offs that justify such preference.

Finally, the robustness analysis showed that ranking stability depends on perturbation magnitude. Across the tested $\pm 5\%$, $\pm 10\%$, and $\pm 15\%$ perturbation levels, the preservation of the full ranking decreased gradually as the magnitude increased, whereas the top-ranked alternative remained comparatively stable. This suggests that the framework is particularly robust for identifying the best alternative, even when the complete ordering becomes more sensitive under stronger perturbations. This emphasis on magnitude-dependent robustness aligns with recent advances in sensitivity and robustness analysis of decision models, which highlight the importance of assessing how different perturbation levels affect rankings and model outputs [17,18]. In recent literature, robustness has become

an important requirement in interpretable and explainable decision analytics, especially because reliable, transparent, and stable explanations are essential for the adoption of AI-based decision-support systems [19,20,24].

Overall, the findings indicate that integrating fuzzy logic and explainability techniques can transform numerical decision-making processes into structured, interpretable knowledge. The proposed framework therefore serves as a bridge between traditional fuzzy decision-making models and hybrid applied AI systems that require both computational rigor and human-centered interpretability.

5.2. Methodological Implications

The proposed framework contributes to the literature by connecting fuzzy representation, multicriteria evaluation, and explainability within a single decision pipeline. From an applied artificial intelligence perspective, this integration can be understood as a hybrid AI system in which fuzzy logic models uncertainty, surrogate modeling approximates the decision function, and SHAP-based analysis decomposes decision outcomes into interpretable criterion-level contributions. While fuzzy MCDM methods and explainability techniques have often been applied separately, their integration has frequently remained partial or heuristic. In contrast, the present framework organizes these components around a decision operator that can be locally approximated and interpreted through surrogate modeling and SHAP decomposition. This responds to the need for explanation-aware analytical pipelines in which ranking, approximation, and interpretation remain conceptually aligned [9,10,12,25–27].

An important implication is that interpretability can be introduced without modifying the original decision mechanism. The surrogate model acts as an explanatory proxy for the fuzzy ranking process, while SHAP values provide a theoretically grounded decomposition of the decision score. This avoids simplifying or linearizing the original model and is consistent with the growing use of surrogate-based strategies in XAI to approximate complex decision functions while preserving interpretability [28]. Methodologically, this is relevant because explanation quality depends not only on transparency, but also on the fidelity of the explanatory proxy with respect to the underlying model [11,29–31]. In applied decision-support contexts, this balance between interpretability and fidelity is essential for producing explanations that are both understandable and reliable.

The comparative surrogate analysis also clarifies the methodological role of model complexity in the proposed framework. Since the surrogate is used only as an explanatory approximation, a simpler model should be preferred whenever it provides comparable fidelity. For this reason, this study compared Random Forest with linear regression and a shallow decision tree. The results showed that the simpler models were useful baselines, but they did not reproduce the FAHP–FTOPSIS decision operator with the same level of fidelity and ranking preservation. Therefore, the Random Forest was retained not because complexity is desirable in itself, but because it provided the strongest local approximation while remaining compatible with SHAP-based criterion-level explanations.

The introduction of SHAP-guided rule extraction further extends the framework beyond numerical attribution. By translating contribution patterns into linguistic IF–THEN rules, the model provides a symbolic representation of decision logic that is more accessible to human stakeholders. This is consistent with research on linguistic fuzzy rule-based systems, where interpretability is achieved through structured rule representations that balance transparency and modeling capacity [21]. It also aligns with recent work showing that explainable MCDM benefits from multi-level explanations and symbolic mechanisms capable of converting model behavior into rule-based justifications that are easier to communicate, validate, and operationalize [12,13,29].

At the same time, the framework speaks to the broader debate in XAI regarding post hoc explanations and inherently interpretable models [32]. Rather than replacing the fuzzy MCDM model with a simplified structure, the proposed approach preserves the FAHP–FTOPSIS mechanism and adds an explanatory layer designed to make its behavior transparent and usable in practice. This hybrid configuration suggests that effective explainability in decision-support systems can be achieved through complementary mechanisms that preserve both fidelity and interpretability [32].

To contextualize the contribution of the proposed framework, Table 11 compares it with representative methodological families in the literature and includes key references associated with each approach. Traditional fuzzy MCDM methods provide robust mechanisms for weighting and ranking alternatives under uncertainty, but they usually offer limited explanation of the internal logic behind the ranking. Sensitivity and robustness analyses improve the assessment of ranking stability, but they do not necessarily translate the decision logic into human-readable explanations. Recent explainable MCDM and hybrid MCDM–machine learning approaches have advanced the use of visual analytics, attribution methods, and predictive models; however, they often remain focused on numerical or graphical explanation outputs. By contrast, the proposed framework combines fuzzy ranking, surrogate-based approximation, SHAP attribution, and linguistic rule extraction in a unified pipeline. Its main advantage is the transformation of fuzzy decision behavior into interpretable IF–THEN rules while preserving the original FAHP–FTOPSIS structure.

Table 11. Comparative positioning of the proposed framework against related methodological approaches, including representative references.

Approach	Representative References	Main Contribution	Explanation Capacity	Main Limitation	Position of the Proposed Framework
Traditional FAHP–FTOPSIS	Buckley [4]; Chen [6]; Mardani et al. [7]; Liu et al. [8]	Handles uncertainty through fuzzy weighting and fuzzy ranking	Limited; mainly based on weights, scores, and rankings	Does not explicitly explain why alternatives obtain specific ranking positions	Preserves the FAHP–FTOPSIS structure but adds surrogate-based SHAP explanations and linguistic rules
Fuzzy MCDM with sensitivity analysis	Borgonovo and Plischke [17]; Gaona et al. [18]	Evaluates ranking stability under changes in weights or inputs	Moderate; explains how variations affect final rankings	Focuses on robustness, not on symbolic explanation of decision logic	Incorporates local perturbation analysis while also extracting interpretable rules
Explainable TOPSIS and visual MCDM approaches	Shi and Yao [12]; Susmaga et al. [19]	Improves transparency through visual or analytical inspection of ranking mechanisms	Moderate to high; useful for understanding weight and aggregation effects	Explanations often remain graphical or numerical	Converts explanation patterns into compact IF–THEN rules
Hybrid MCDM–machine learning approaches	Abdulla and Baryannis [20]; Guidotti et al. [25]; Bodria et al. [26]	Combines multicriteria evaluation with predictive or classification models	Variable; depends on the interpretability of the machine learning component	May replace or obscure the original MCDM logic	Uses the surrogate only as an explanatory approximation, not as a replacement of the fuzzy decision operator
SHAP-based explainability methods	Lundberg et al. [11]; Khalyasmaa et al. [13]	Provides local and global feature-attribution explanations	High for numerical attribution	Attribution values may be difficult for non-technical users to operationalize	Uses SHAP values as an intermediate layer for linguistic rule extraction
Linguistic fuzzy rule-based systems	Herrera [21]; Khalyasmaa et al. [13]	Provides interpretable symbolic rules	High when the rule base is compact	Rule generation may be detached from a formal MCDM ranking mechanism	Links rule extraction directly to the fuzzy MCDM decision operator through SHAP-guided patterns

This comparison also highlights the scope and limitations of the proposed approach. First, its explanatory capacity depends on surrogate fidelity; if the surrogate does not adequately approximate the fuzzy decision operator, the resulting SHAP values and rules may be unreliable. Second, discretizing SHAP values into linguistic categories improves interpretability but may reduce numerical precision and introduce sensitivity to threshold choices. To address this issue, the analysis tested alternative percentile schemes for

SHAP discretization, comparing the baseline 25th/75th configuration with 20th/80th and 30th/70th thresholds. The results showed that the dominant rule patterns remained stable across the tested configurations, supporting the robustness of the extracted rules while confirming the need to report threshold sensitivity explicitly.

In addition, the rule evaluation compared the SHAP-guided rules against a majority-class baseline and reported bootstrapped confidence intervals for coverage and consistency. This strengthens the interpretability claim because the rule set is not evaluated only through point estimates, but against a trivial explanatory benchmark and with uncertainty measures. Third, rule extraction may become more complex as the number of alternatives, criteria, and perturbation scenarios increases. Therefore, the proposed framework should be understood as a contribution to explainable fuzzy decision support, rather than as a replacement for existing fuzzy MCDM, sensitivity analysis, or interpretable machine learning methods.

5.3. Practical Implications

From an applied perspective, the proposed approach offers several advantages for decision-making in organizational and engineering contexts. First, it enhances transparency by explaining why certain alternatives are preferred, which is particularly relevant when decisions must be justified to stakeholders, audited, or negotiated. This is especially important in supplier evaluation and related selection problems, where limited interpretability can hinder the adoption of advanced analytical models. In this sense, explainable AI techniques contribute to trust, accountability, and user acceptance by making the reasoning behind decisions more accessible and verifiable. This is particularly relevant in supplier selection contexts, where transparency and sustainability considerations increasingly influence procurement decisions [20,33].

Compared with traditional fuzzy MCDM applications, the proposed framework changes the practical use of the model from a ranking-only tool to a diagnostic and communicative decision-support mechanism. In conventional fuzzy MCDM, decision-makers typically obtain a final score or ranking and must infer the rationale from weights and performance matrices. In contrast, the proposed approach identifies the criterion-level contributions behind each ranking position, making it possible to explain which strengths compensate for weaknesses and which trade-offs justify the final decision.

Second, the combination of global importance measures, local explanations, and linguistic rules allows decision-makers to analyze the problem at different levels of abstraction. Global SHAP values identify the most influential criteria, local explanations clarify specific ranking outcomes, and rule-based representations translate these insights into actionable decision guidelines. This multi-layered interpretability enables decision-makers not only to justify outcomes, but also to explore alternative scenarios and validate decisions under different conditions.

The extracted IF–THEN rules can be used as operational decision heuristics. For example, a rule such as “IF quality is high AND delivery is high THEN score is high” provides a concise justification of why a supplier is favored, while rules involving high cost and high risk identify unfavorable profiles that should be monitored or avoided. In practice, these rules can support supplier screening, communication of results to non-technical stakeholders, auditing of selection decisions, and the definition of minimum acceptable performance profiles for future evaluations. The practical usefulness of these rules is strengthened by the fact that their coverage and consistency were evaluated against a majority-class baseline, with bootstrapped confidence intervals and threshold sensitivity checks. Thus, the rule layer transforms the model from a purely computational ranking procedure into a transparent knowledge-support tool for managerial decision-making.

Third, the robustness of the framework supports its use in environments characterized by uncertainty and variability. The results indicate that ranking stability is magnitude-dependent: the complete ranking becomes more sensitive as perturbation magnitude increases, whereas the top-ranked alternative remains comparatively stable across the tested perturbation levels. This distinction is practically relevant because decision-makers are often more concerned with the reliability of the best alternative than with the exact preservation of every ranking position. Real-world decision problems, such as supplier selection or project evaluation, often involve imprecise, incomplete, or subjective information. In this context, combining robustness and interpretability is essential because unstable rankings or opaque explanations can reduce user trust and hinder adoption [8,20].

Finally, the framework facilitates the integration of advanced analytics into practical decision-support systems. By embedding explainability directly into the decision workflow, rather than treating it as an external add-on, the approach reduces the gap between model development and real-world implementation. This is particularly relevant in organizational contexts involving multiple stakeholders, regulatory constraints, and the need for traceability. Overall, the proposed framework is especially suitable for applications such as supplier selection, project prioritization, and risk evaluation, where decisions must balance multiple criteria while remaining transparent, robust, and defensible. In such settings, hybrid combinations of MCDM and explainable AI can enhance both analytical rigor and stakeholder acceptance [20,34].

For larger-scale decision-support systems, implementation should carefully address criterion hierarchy design, perturbation sampling, surrogate-model selection, validation strategy, and rule pruning so that explanation outputs remain interpretable and operationally useful.

5.4. Limitations and Future Research

Despite its contributions, the proposed framework presents several limitations that define the scope within which the results should be interpreted and indicate directions for future research.

First, the empirical validation is based on a small illustrative supplier-selection case involving six original alternatives and three experts. This configuration is useful for demonstrating the operational logic of the proposed framework, but it does not provide a sufficient empirical basis for population-level generalization. Although local perturbations were generated around the original supplier profiles, these instances remain dependent on the six parent alternatives and should therefore be interpreted only as a basis for local approximation, robustness testing, and explanation analysis. To address this dependence, the validation strategy uses leave-one-supplier-out cross-validation rather than a random train–test split. In addition, a diverse synthetic benchmark was incorporated to examine the framework under broader decision configurations. Nevertheless, future research should validate the approach using real supplier datasets involving larger numbers of alternatives, multiple procurement contexts, and external performance indicators. Real-world datasets would allow for stronger conclusions regarding predictive generalization, explanation stability, and practical decision-support performance. Although robustness was evaluated across multiple perturbation magnitudes, the analysis remains local and bounded around the original supplier profiles. Larger deviations, correlated perturbations among criteria, changes in criteria weights, or alternative uncertainty structures may produce different stability patterns. Future studies should therefore extend robustness analysis to include weight perturbations, scenario-based stress tests, and global sensitivity analysis.

Second, the framework may face scalability challenges as the dimensionality of the decision problem increases. Larger applications would require more extensive fuzzy evalu-

ation matrices, more complex pairwise comparison structures, larger perturbation datasets for surrogate training, and more complex SHAP attribution patterns. The linguistic rule extraction stage may also generate many candidate rules, increasing the need for pruning, redundancy control, and interpretability management. Future applications could address these challenges through hierarchical or clustered criteria structures, optimized perturbation sampling, cross-validation, parallel computation, automated hyperparameter tuning, minimum coverage thresholds, and rule aggregation procedures, drawing on recent advances in multi-stage group decision-making under uncertain preference information [35]. Although this study included a perturbation-size convergence analysis and found that 100 perturbations per alternative were sufficient for the present case, this number should not be treated as a universal default. Larger, more heterogeneous, or higher-dimensional decision problems may require additional perturbation instances to stabilize surrogate fidelity, SHAP attributions, and rule extraction. Future studies should therefore report perturbation-size convergence checks when applying the framework to new contexts.

Third, the use of triangular fuzzy numbers, while computationally efficient and widely adopted, represents a simplified model of uncertainty. More advanced fuzzy representations, such as interval type-2 fuzzy sets, hesitant fuzzy sets, or intuitionistic fuzzy sets, could better capture ambiguity, hesitation, or disagreement among experts. However, these extensions would increase computational complexity and may affect interpretability. Future research should therefore examine the trade-off between representational richness, computational cost, and explainability.

Fourth, the explanation layer depends on the fidelity of the surrogate model. Although the analysis compared Random Forest with simpler surrogate baselines, including linear regression and a shallow decision tree, the reliability of SHAP values and extracted rules still depends on how well the selected surrogate approximates the original FAHP–FTOPSIS decision operator. In this study, the Random Forest provided the strongest local fidelity and was therefore retained for the SHAP-based explanation stage. However, this result should not be generalized as a universal preference for Random Forest surrogates. In other decision contexts, simpler models may achieve comparable fidelity and should be preferred when they provide sufficient approximation accuracy with greater transparency. Future studies should evaluate explanation stability across additional model classes and incorporate broader fidelity metrics for explainable multicriteria decision-making. Moreover, the FAHP–SHAP alignment analysis was based on six criteria; therefore, although the observed Spearman correlation supports internal coherence in the present case study, future applications with larger criteria sets should reassess this alignment and report uncertainty measures such as confidence intervals or resampling-based stability indicators.

Fifth, the rule extraction process relies on discretizing continuous SHAP values into qualitative contribution levels. This transformation improves interpretability and facilitates communication with stakeholders, but it may reduce numerical precision and introduce sensitivity to threshold choices. To address this issue, this study incorporated a threshold sensitivity analysis comparing 20th/80th, 25th/75th, and 30th/70th percentile schemes. The results showed that the dominant rule patterns remained stable across the tested configurations. In addition, the analysis incorporated a majority-class baseline and bootstrapped confidence intervals for rule-quality indicators. Nevertheless, future studies should explore automated discretization methods, adaptive threshold selection, entropy-based binning, formal rule-stability measures, and additional baselines such as random rule sets, decision-tree-derived rules, or association-rule mining approaches in larger and more heterogeneous decision spaces.

Sixth, the framework has not yet been evaluated through user studies. The extracted rules are designed to be interpretable, but the present study does not empirically assess

how decision-makers understand, trust, or use these explanations in practice. Future research should conduct user-centered evaluations with managers, domain experts, and stakeholders to assess the usefulness of explanations, usability, cognitive load, perceived transparency, and decision confidence.

Finally, future work could extend the framework toward dynamic and interactive decision-support systems. Incorporating time-dependent criteria would allow the model to address evolving decision environments, such as supply chain management or risk assessment. Embedding the framework in interactive interfaces would also allow decision-makers to explore ranking scenarios, conduct sensitivity analyses, and view explanation outputs in real time, thereby enhancing usability and practical impact.

In summary, the proposed framework should be interpreted as a methodological contribution and proof-of-concept for explainable fuzzy MCDM, rather than as a fully generalizable decision-support system. Future research should extend the framework through larger-scale empirical applications, systematic synthetic benchmarks, comparative surrogate modeling, adaptive rule-extraction procedures, user-centered validation, and more expressive fuzzy-uncertainty representations.

6. Conclusions

This study proposed an explainable fuzzy multi-criteria decision-making framework that integrates Fuzzy AHP, Fuzzy TOPSIS, surrogate modeling, SHAP-based explanation, and linguistic rule extraction within a unified decision architecture. The objective was to address a key limitation of conventional fuzzy MCDM methods: their limited capacity to provide transparent and interpretable explanations of ranking outcomes under uncertainty.

The results show that the proposed framework can generate discriminative rankings while supporting interpretability at both global and local levels. The quantitative alignment between FAHP-derived weights and SHAP-based global importance values indicates that the explanatory layer remains consistent with the decision model's normative structure. In addition, the surrogate-model comparison showed that Random Forest provided the strongest local approximation among the evaluated candidate models, supporting its use as an explanatory proxy for the FAHP–FTOPSIS decision operator. Importantly, this surrogate was not used to replace the fuzzy MCDM procedure, but to decompose its ranking behavior into criterion-level contributions.

This study also strengthened the evaluation of the explanation layer. The use of leave-one-supplier-out cross-validation reduced information leakage associated with dependent perturbations, while the synthetic benchmark extended the assessment beyond the immediate neighborhood of the six original supplier profiles. The perturbation-size convergence analysis supported using 100 perturbations per alternative in the present case, and the robustness analysis showed that ranking stability depends on the perturbation magnitude. In particular, the complete ranking became more sensitive as perturbation magnitude increased, whereas the top-ranked alternative remained comparatively stable across the tested perturbation levels.

From a methodological perspective, the main contribution of this study lies in establishing a coherent integration between fuzzy multicriteria evaluation and explainable artificial intelligence. By incorporating a surrogate-based explanation layer and SHAP-guided rule extraction, the framework connects numerical decision modeling with symbolic knowledge representation without altering the original FAHP–FTOPSIS decision operator. The extracted linguistic rules transform numerical attribution patterns into explicit and human-readable decision logic. Their stability was further examined through threshold sensitivity analysis, majority-class baseline comparison, and bootstrapped confidence intervals for rule-quality indicators.

However, this study is subject to limitations related to the scale of the empirical validation, the use of triangular fuzzy numbers, the dependence of local perturbations on the original supplier profiles, and the reliance on surrogate modeling. The six original supplier alternatives should be understood as an illustrative decision scenario rather than as a statistically representative empirical sample. Similarly, the perturbation-based dataset supports the assessment of local surrogate fidelity, explanation stability, and robustness within a bounded decision space, but it does not provide population-level predictive validation. Broader validation with larger real-world supplier datasets, additional synthetic benchmarks, more heterogeneous decision contexts, and user-centered evaluations is required before making general claims about performance across different applications.

In conclusion, this work shows that fuzzy multicriteria decision-making and explainable artificial intelligence can be integrated into a coherent framework that preserves uncertainty handling while improving transparency, traceability, and interpretability. The proposed approach should be understood as a methodological contribution and proof of concept for explainable fuzzy MCDM, rather than as a fully generalizable decision-support system. Future research should extend the framework through larger-scale empirical applications, alternative fuzzy representations, comparative surrogate modeling, adaptive rule-extraction procedures, global sensitivity analysis, and interactive decision-support implementations.

Author Contributions: Conceptualization, J.A.R.-F. and A.S.-R.; software, Y.F.-O. and J.A.R.-F.; methodology, G.G.-V., A.C.-G. and R.P.-C.; validation, G.G.-V. and R.P.-C.; formal analysis, A.S.-R., A.C.-G. and J.A.R.-F.; investigation, G.G.-V., A.S.-R., R.P.-C. and J.A.R.-F.; data curation, R.P.-C., Y.F.-O. and A.S.-R.; writing—original draft preparation, J.A.R.-F. and A.S.-R.; writing—review and editing, A.S.-R.; visualization, A.C.-G., Y.F.-O. and R.P.-C.; supervision, G.G.-V.; project administration, R.P.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The dataset used in this research was generated for methodological validation purposes and is not derived from proprietary or confidential sources. All procedures, model specifications, and computational steps required to reproduce the results are fully described within this article.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI) (GPT-5.5 Thinking) only for language editing, grammar correction, wording refinement, and improvement in academic writing style. The tool was not used to generate this article's scientific content, formulate the methodology, produce or analyze results, derive conclusions, or create the proposed decision framework. All conceptual development, methodological design, mathematical formulation, data processing, interpretation of findings, and final intellectual content were carried out, reviewed, and approved by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Triantaphyllou, E. *Multi-Criteria Decision Making Methods: A Comparative Study*; Springer: New York, NY, USA, 2000. [CrossRef]
2. Zavadskas, E.K.; Turskis, Z.; Kildienė, S. State of art surveys of overviews on MCDM/MADM methods. *Technol. Econ. Dev. Econ.* **2014**, *20*, 165–179. [CrossRef]
3. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
4. Buckley, J.J. Fuzzy hierarchical analysis. *Fuzzy Sets Syst.* **1985**, *17*, 233–247. [CrossRef]

5. Chang, D.-Y. Applications of the Extent Analysis Method on Fuzzy AHP. *Eur. J. Oper. Res.* **1996**, *95*, 649–655. [CrossRef]
6. Chen, C.-T. Extensions of the TOPSIS for group decision-making under fuzzy environment. *Fuzzy Sets Syst.* **2000**, *114*, 1–9. [CrossRef]
7. Mardani, A.; Jusoh, A.; Zavadskas, E.K. Fuzzy Multiple Criteria Decision-Making Techniques and Applications—Two Decades Review from 1994 to 2014. *Expert Syst. Appl.* **2015**, *42*, 4126–4148. [CrossRef]
8. Liu, Y.; Eckert, C.M.; Earl, C. A review of fuzzy AHP methods for decision-making with subjective judgements. *Expert Syst. Appl.* **2020**, *161*, 113738. [CrossRef]
9. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
10. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
11. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
12. Shi, C.; Yao, Y. Explainable multi-criteria decision-making: A three-way decision perspective. *Int. J. Approx. Reason.* **2025**, *187*, 109528. [CrossRef]
13. Khalyasmaa, A.I.; Matrenin, P.V.; Eroshenko, S.A. Interpretable Diagnostics with SHAP-Rule: Fuzzy Linguistic Explanations from SHAP Values. *Mathematics* **2025**, *13*, 3355. [CrossRef]
14. Saaty, T.L. *The Analytic Hierarchy Process*; McGraw-Hill: New York, NY, USA, 1980.
15. Hwang, C.L.; Yoon, K. *Multiple Attribute Decision Making: Methods and Applications*; Springer: Berlin, Germany, 1981.
16. Lu, Z.; Wang, W.; Guo, T.; Li, Y.; Wang, F. Decoding urban policies: NLP-driven concise explanations. *Environ. Plan. B Urban Anal. City Sci.* **2025**, *53*, 125–142. [CrossRef]
17. Borgonovo, E.; Plischke, E. Sensitivity analysis: A review of recent advances. *Eur. J. Oper. Res.* **2016**, *248*, 869–887. [CrossRef]
18. Gaona, À.; Guisasola, A.; Baeza, J.A. An integrated TOPSIS framework with Full-Range Weight Sensitivity Analysis for robust decision analysis. *Decis. Anal. J.* **2025**, *17*, 100642. [CrossRef]
19. Susmaga, R.; Szczęch, I.; Brzezinski, D. Towards explainable TOPSIS: Visual insights into the effects of weights and aggregations on rankings. *Appl. Soft Comput.* **2024**, *153*, 111279. [CrossRef]
20. Abdulla, A.; Baryannis, G. A Hybrid Multi-Criteria Decision-Making and Machine Learning Approach for Explainable Supplier Selection. *Supply Chain Anal.* **2024**, *7*, 100074. [CrossRef]
21. Gacto, M.J.; Alcalá, R.; Herrera, F. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Inf. Sci.* **2011**, *181*, 4340–4360. [CrossRef]
22. Behzadian, M.; Otaghsara, S.K.; Yazdani, M.; Ignatius, J. A State-of the-Art Survey of TOPSIS Applications. *Expert Syst. Appl.* **2012**, *39*, 13051–13069. [CrossRef]
23. Opricovic, S.; Tzeng, G.-H. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *Eur. J. Oper. Res.* **2004**, *156*, 445–455. [CrossRef]
24. Celik, E.; Gul, M.; Aydin, N.; Gumus, A.T.; Guneri, A.F. A Comprehensive Review of Multi Criteria Decision Making Approaches Based on Interval Type-2 Fuzzy Sets. *Knowl.-Based Syst.* **2015**, *85*, 329–341. [CrossRef]
25. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2019**, *51*, 93. [CrossRef]
26. Bodria, F.; Giannotti, F.; Guidotti, R.; Naretto, F.; Pedreschi, D.; Rinzivillo, S. Benchmarking and survey of explanation methods for black box models. *Data Min. Knowl. Discov.* **2023**, *37*, 1719–1778. [CrossRef]
27. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput. Surv.* **2023**, *55*, 295. [CrossRef]
28. Monteiro, W.R.; Reynoso-Meza, G. A multi-objective optimization design to generate surrogate machine learning models in explainable artificial intelligence applications. *EURO J. Decis. Process.* **2023**, *11*, 100040. [CrossRef]
29. Pensa, R.G.; Crombach, A.; Peignier, S.; Rigotti, C. Explaining Random Forest and XGBoost with Shallow Decision Trees by Co-Clustering Feature Importance. *Mach. Learn.* **2025**, *114*, 287. [CrossRef]
30. Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665. [CrossRef]
31. Pelegrina, G.D.; Duarte, L.T.; Grabisch, M. A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artif. Intell.* **2023**, *325*, 104014. [CrossRef]
32. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]

33. Schramm, V.B.; Cabral, L.P.B.; Schramm, F. Approaches for supporting sustainable supplier selection—A literature review. *J. Clean. Prod.* **2020**, *273*, 123089. [CrossRef]
34. Özkurt, C. Explainable AI for interpreting decision processes in multi-criteria decision making for robotic systems. *Knowl.-Based Syst.* **2026**, *340*, 115700. [CrossRef]
35. Tu, Y.; Ma, Z.; Liu, J.; Zhou, X.; Lev, B. Multi-Stage Multi-Criteria Group Decision Making Method Based on Hierarchical Criteria Interaction and Trust Rejection Threshold under Uncertain Preference Information. *Inf. Sci.* **2023**, *647*, 119509. [CrossRef]