Interpretable and Uncertainty-Aware by Construction Decisions with Fuzzy Theory

Diego Perez-Palacin[1,2(✉)], Vincenzo Scotti[2], Alireza Maleki[2], Vincenzo Grassi[3], and Raffaela Mirandola[2]

[1] Linnaeus University, Växjö, Sweden
diego.perez@lnu.se
[2] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{vincenzo.scotti,alireza.maleki,raffaela.mirandola}@kit.edu
[3] University of Roma Tor Vergata, Roma, Italy
vincenzo.grassi@uniroma2.it

Abstract. Autonomous systems increasingly use Machine-Learning (ML) components to perceive their environment and guide the decision-making process. However, the confidence scores of these components are often miscalibrated and fail to represent true predictive reliability. While existing calibration methods reduce this issue, they still compress predictions and uncertainty into a single value, losing significant information. In this work, we propose a method to have decision-making "uncertainty-aware by construction" by applying fuzzy theory to represent and manipulate uncertainty in ML confidence scores. Our approach extends a standard perception-decision pipeline with two new modules: an Estimator Performance component that collects diverse calibration and error indicators, and a Fuzzifier that transforms these indicators, together with the raw confidence score, into a fuzzy number encoding second-order uncertainty. This fuzzy confidence is then consumed by a fuzzy decision-maker capable of adapting its interpretation of uncertainty to the system's current goals. Experiments on a classification dataset show that fuzzy confidence retains uncertainty information discarded by traditional calibration, enabling better-aligned decisions under changing goals.

1 Introduction

Autonomous systems increasingly rely on decision-making under uncertainty, where it is essential to provide not only predictions but also reliable measures of confidence in the correctness of those predictions. Equally important, both predictions and confidence must be *interpretable*, enabling system designers and operators to understand, trust, and act upon them. This requirement has grown in importance as perception and estimation components are increasingly based on Machine Learning (ML).

ML models typically output predictions along with confidence scores intended to reflect predictive reliability. However, extensive work has shown that such

© The Author(s) 2027
T. Margaria and B. Steffen (Eds.): ISoLA 2026, LNCS 17144, pp. 115–135, 2027.
https://doi.org/10.1007/978-3-032-40121-2_7

scores are often miscalibrated, particularly in modern models, and do not correspond to true *a posteriori* correctness probabilities [7,10]. Both overconfidence and underconfidence undermine downstream decision-making, which assumes that confidence estimates are meaningful and reliable.

To address this issue, numerous calibration techniques have been proposed, including post-hoc methods such as Platt scaling, isotonic regression, and temperature scaling [10]. Despite progress, calibration remains an open problem. Most approaches reduce uncertainty to a scalar confidence value. From an interpretability standpoint, this simplification obscures underlying uncertainty and often proves insufficient for decision-making in dynamic settings [12], particularly in the presence of irreducible (aleatoric) uncertainty.

An alternative is to augment predictions with uncertainty bounds, such as confidence intervals. While these acknowledge uncertainty, they offer limited interpretability for complex decision processes and often rely on restrictive assumptions or boundary-based reasoning.

Probabilistic approaches, including Bayesian modeling, address these limitations by representing outputs as full probability distributions [19,25]. Although primarily explored in regression, this idea extends to confidence estimation in classification, capturing richer forms of uncertainty such as *second-order uncertainty*. However, such methods typically incur high computational and modeling costs and can be difficult to interpret, limiting their practical applicability in autonomous systems [19].

This motivates the need for a practical middle ground between full probabilistic modeling and overly simplified scalar confidence values, with a strong emphasis on interpretability. In this context, hybrid approaches combining sub-symbolic ML components with symbolic reasoning mechanisms have received increasing attention [3]. Sub-symbolic models excel at perception and pattern recognition, while symbolic approaches offer transparency, interpretability, and structured reasoning. Fuzzy theory naturally occupies a central position in this spectrum, as it provides symbolic representations that are both mathematically grounded and linguistically interpretable, while remaining tolerant to vagueness and imprecision [14]. Moreover, fuzzy systems have long been recognized for their interpretability and explainability [18], making them particularly suitable for decision-making in autonomous and safety-related systems. By integrating fuzzy theory with ML-based confidence estimation, it becomes possible to combine the expressive power of sub-symbolic prediction with the interpretability and transparency of symbolic reasoning, contributing to the broader goals of Explainable Artificial Intelligence.

In this work, we leverage fuzzy theory to achieve uncertainty-aware behavior by construction, while preserving interpretability. We introduce UAFuCo, Uncertainty-Aware Fuzzy Confidence, which extends a standard estimation-decision pipeline with two components. An *Estimator Performance* module aggregates heterogeneous calibration and error indicators characterizing the behavior of the estimator in an interpretable manner. A *Fuzzifier* then com-

bines these indicators with the raw confidence score to construct a fuzzy number encoding second-order uncertainty.

The resulting fuzzy confidence is consumed by a fuzzy decision-maker that adapts its interpretation of uncertainty to the system's current goals through interpretable rule sets. This property makes the proposed approach particularly suitable for self-adaptive systems with evolving goals and preferences [26], where transparency of decision-making logic is essential. From a software engineering perspective, the method can be applied as a wrapper around existing estimators, allowing them to be used out of the box without requiring access to their internal structure, while enhancing their interpretability. Moreover, the approach is estimator-agnostic and can be applied to ML predictors, sensor fusion mechanisms, and other estimation components.

Experiments on a classification dataset show that fuzzy confidence representations preserve uncertainty information typically lost in standard calibration, while maintaining high interpretability. These findings position fuzzy theory as a practical and expressive alternative for uncertainty representation in autonomous decision-making systems when transparency and explainability are essential.

The remainder of the paper is organized as follows. Section 2 summarizes the background and the related work. Section 3 introduces a motivating scenario for the proposed approach presented in Sect. 4. Section 5 illustrates the evaluation performed on an exemplary dataset and Sect. 6 concludes the paper.

2 Background and Related Work

Uncertainty quantification and confidence estimation in Machine Learning have attracted extensive attention in recent years, resulting in a fragmented and heterogeneous body of literature. A recent survey by Del Corso et al. [4] highlights the lack of conceptual uniformity and the coexistence of multiple uncertainty formalisms, including probabilistic, evidential, and fuzzy approaches. In this section, we review the most relevant background on confidence calibration (Sect. 2.1), uncertainty estimation (Sect. 2.2), and fuzzy representations for decision-making (Sect. 2.3).

2.1 Confidence Calibration and Evaluation

Confidence calibration addresses the mismatch between predicted confidence values and the empirical probability of correctness. Classical *post-hoc* calibration techniques include *Platt scaling* [21], *isotonic regression* [29], and *temperature scaling* [5]. These methods transform model outputs without modifying their internal structure and are widely used due to their simplicity and effectiveness. Comparative studies have demonstrated their applicability across a wide range of learning algorithms [20], as well as in domain-specific scenarios such as medical imaging [22]. Despite their success, calibration approaches reduce uncertainty to a single scalar value and assume that a unique "true" probability of correctness exists. This assumption is problematic in the presence of irreducible uncertainty, motivating richer representations of confidence.

2.2 Predictive Uncertainty in Machine Learning

Beyond post-hoc calibration, several approaches integrate uncertainty estimation directly into the prediction process. In regression, uncertainty is often expressed through confidence intervals or predictive distributions [1,25], while similar ideas are less explored in classification. Probabilistic techniques such as Bayesian Neural Networks (BNNs) [19] and Monte Carlo dropout [7] approximate predictive distributions but often incur significant computational costs.

In particular, BNNs provide a principled framework for uncertainty estimation by placing distributions over model parameters and deriving predictive distributions. While this allows for modeling epistemic uncertainty, practical implementations rely on approximate inference due to the intractability of the computations, which leads to the introduction of assumptions that may limit expressiveness. Moreover, BNNs typically require multiple stochastic forward passes during inference, increasing computational cost and latency. Finally, the resulting uncertainty representations, often expressed as high-dimensional distributions, are not always easily interpretable in downstream decision-making. These limitations motivate alternative approaches that preserve uncertainty information while remaining computationally efficient and interpretable.

Moreover, the relationship between predictive uncertainty and actual error is not always reliable, particularly in safety-critical tasks [24]. These limitations motivate approaches that preserve richer uncertainty information without requiring full probabilistic modeling.

2.3 Fuzzy Representations for Uncertainty and Decision-Making

Fuzzy theory provides a well-established framework for representing imprecision through graded membership functions [28]. It has been widely applied in control, uncertainty modeling, and multi-criteria decision-making. *Fuzzy numbers*, in particular, enable the representation of uncertain numerical quantities and support operations such as aggregation and defuzzification [11,15,27]. They have also been successfully applied in industrial and trust-aware decision-making contexts [8,17] and in the assessment of the fulfillment of uncertain goals [9].

The construction of fuzzy confidence values from heterogeneous uncertainty indicators is conceptually related to fuzzy sensor fusion, where multiple information sources are combined into a unified representation [2,6,23]. In contrast to classical fusion settings, uncertainty indicators may originate from heterogeneous domains, making fuzzy aggregation particularly suitable.

Formally, fuzzy sets and fuzzy numbers are defined as follows. Let X be a set and $\mathcal{F}(X) = \{A | A : X \to [0,1]\}$ the set of all functions from X to interval $[0,1]$. A fuzzy set defined over X is characterized by the *membership function* $A \in \mathcal{F}(X)$, where, given $x \in X$, $0 \leq A(x) \leq 1$ represents the *membership degree* of x to A. A fuzzy set $A \in \mathcal{F}(X)$ defines its *support* as $\text{supp}(A) = \{x \in X \mid A(x) > 0\}$, and its *core* as $\text{core}(A) = \{x \in X \mid A(x) = 1\}$. A *normal* fuzzy set A holds that $core(A) \neq \emptyset$. The α-cut of a fuzzy set A is a crisp set defined as $A_\alpha = \{x \in X \mid A(x) \geq \alpha\}$.

Fuzzy numbers are able to represent vagueness in the value of a number. A fuzzy number is a fuzzy set A defined on the set of real numbers $\mathbb{R}$ and that holds the following properties [14]: *(i)* A must be a normal fuzzy set *(ii)* A_α must be a closed interval for all $\alpha \in (0, 1]$ *(iii)* the support of A must be bounded.

In the context of confidence estimation, a fuzzy number B that generalizes scalar confidence values in $[0, 1]$ by encoding second-order uncertainty must hold $supp(B) \subseteq [0, 1]$.

In summary, existing work provides a rich set of methods for estimating confidence and uncertainty. However, most approaches either collapse uncertainty into a single scalar or require complex probabilistic modeling. Fuzzy theory offers a complementary representation that preserves uncertainty information while remaining interpretable and compatible with symbolic decision-making mechanisms.

3 Motivating Scenario

Consider an autonomous unmanned aerial vehicle (UAV) deployed for accident response and rescue operations in mountainous environments, such as dense forests or cave entrances. In such settings, mobile connectivity is often unavailable, preventing injured individuals from calling for help or transmitting their location. The initial mission phase, therefore, consists of locating the victim. Once the victim is found, the mission transitions to assisting in the rescue process. Although the same UAV is used throughout the mission, its operational goals differ across these phases.

During the search phase (P_1), the UAV explores a complex and partially unknown environment, navigating through obstacles such as trees, narrow passages, and possibly windy conditions. While minor collisions (e.g., brushing against branches) or slight changes in the trajectory (e.g., due to the wind) may occur, the UAV is assumed to tolerate these events without mission failure; only severe crashes are considered harmful. Using onboard sensors and ML-based predictors, the UAV classifies candidate trajectories as *safe* or *unsafe*, together with an associated confidence value. In this phase, the UAV may deliberately select riskier trajectories if doing so improves coverage efficiency and increases the likelihood of locating the victim.

After locating the victim, the mission enters the assistance phase (P_2), where the trajectory safety becomes stricter. In P_2, the UAV must guide the person back toward a safe location (e.g., a rescue point). In this case, the UAV must remain within the victim's line of sight, follow a trajectory that is easy to track, and avoid maneuvers that could confuse or endanger the person. Stability, predictability, and continuous visibility become critical constraints, and uncertainty in navigation decisions must be interpreted conservatively.

This scenario illustrates how decision-making requirements change across mission phases. Consider two candidate trajectories. Trajectory t_1 is classified as safe with confidence 0.9 by a perfectly calibrated estimator, whereas trajectory t_2 is also classified as safe with confidence 0.9 by an estimator subject to

a ±10% error margin. In this case, t_1 may be acceptable in both $P1$ and $P2$, whereas t_2 is unsuitable in the risk-averse $P2$. If the estimators output only a scalar confidence score of 0.9, the downstream decision-maker lacks the necessary information to distinguish between these two cases, since relevant information about the trajectory's uncertainty is obscured. Furthermore, this simplification severely limits the interpretability of failed decisions; when a single scalar serves as the sole basis for reasoning, system designers and operators cannot effectively diagnose whether a failure stemmed from estimator unreliability, environmental noise, or model overconfidence.

4 Uncertainty-Aware Fuzzy Confidence UAFuCo

Figure 1 illustrates software pipelines for uncertainty-aware decision-making. Figure 1a depicts the conventional estimation–decision workflow in which an estimator produces a prediction and an associated crisp confidence score that a calibration module may adjust before being consumed by the decision-maker. Figure 1b shows the proposed extension, in which, leveraging fuzzy theory, uncertainty is explicitly represented and propagated rather than collapsing into a single scalar value.

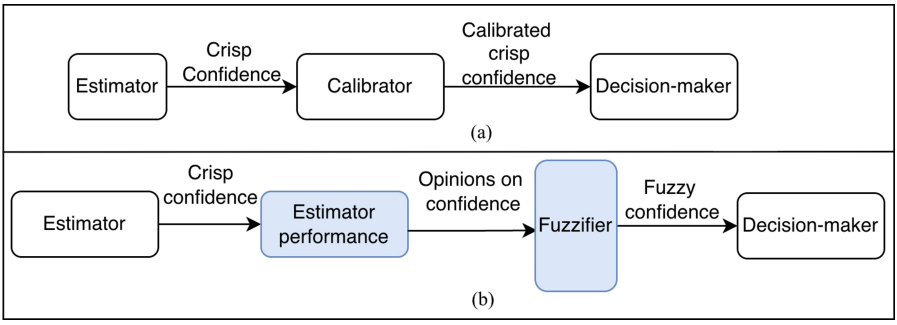


Fig. 1. Estimation-decision pipelines: (a) usual post-hoc calibration (b) UAFuCo.

In the proposed pipeline, the *Estimator* remains unchanged and may correspond to any predictive component, including ML classifiers or sensor-based estimators. The applicability of the approach is intended for *Decision-makers* that are capable of reasoning over uncertainty expressed as fuzzy sets. Two modules are added or reinterpreted: an *Estimator Performance* module and a new *Fuzzifier* component. Unlike traditional calibration approaches, the Estimator Performance module does not modify the Estimator's confidence score. Instead, it gathers heterogeneous indicators related to the reliability of the confidence estimation and its prediction error. The Fuzzifier then combines these indicators to construct a fuzzy representation of confidence, which the decision-maker subsequently consumes.

This design explicitly separates prediction, uncertainty characterization, and goal-dependent decision making. As a result, it facilitates an uncertainty-aware process by preserving the prediction's confidence uncertainty for as long as possible in the pipeline. Furthermore, changes in the system's goals can be accommodated by modifying only the decision-making component that interprets confidence, without requiring changes to the prediction or uncertainty characterization stages. Moreover, uncertainty is represented using a structured formalism grounded in established theory, rather than an *ad hoc* representation, thereby facilitating the engineering of interpretable and reusable decision-makers.

The next subsections present each novel component and discuss the suitability of the approach, comparing it with alternative solutions.

4.1 Estimator Performance

The *Estimator Performance* module characterizes the reliability of the confidence score produced by an estimator. Its purpose is not to select or apply a calibration method, but to expose multiple complementary views of how the estimator confidence behaves with respect to previously observed prediction outcomes.

In traditional model calibration pipelines, a single calibration technique is chosen and applied to transform the estimator confidence score into a corrected scalar value. In contrast, the proposed Estimator Performance module generates multiple *opinions* about the value and reliability of the crisp confidence score by applying several calibration and error-estimation methods in parallel. Each value produced by this module is interpreted as an *opinion* expressing partial information about the reliability of the predicted confidence. These possibly heterogeneous and conflicting opinions are explicitly retained and forwarded to the Fuzzifier, where they serve as uncertainty indicators for constructing a fuzzy representation of confidence. This interpretation aligns naturally with fuzzy modeling, where multiple partial and uncertain assessments are aggregated rather than reconciled into a single precise value.

Abstract Estimator Performance: The Estimator Performance uses

- A set $\mathcal{C}$ of calibration techniques that transform each confidence score produced by the Estimator.
- A set Err of error estimation methods. Each method calculates a prediction error and a miscalibration metric that quantifies the empirical behavior of a calibrator. Each combination (c, e) where $c \in \mathcal{C} \cup \{Uncalibrated\}$ and $e \in Err$ creates an interval which corresponds to an opinion of where the true confidence value lies.

This module generates $(|\mathcal{C}|+1)\cdot(|Err|+1)$ opinions for each confidence score received by the Estimator, namely the $|Err|$ error values for each of the elements in $\mathcal{C} \cup \{Uncalibrated\}$, the crisp confidence scores predicted by each of the calibrators, and the propagated uncalibrated score received from the Estimator.

A Concrete Estimator Performance: In our implementation, we employ widely used post hoc calibration methods and error estimates. The calibration methods are the Platt scaling, temperature scaling, and isotonic regression. Let $\mathcal{C} = \{Pl, Ts, Ir\}$ denote the set of calibration methods considered. Given the confidence score $s \in [0, 1]$ produced by the Estimator, we denote by $Pl(s) = s^{Pl}, \quad Ts(s) = s^{Ts}, \quad Ir(s) = s^{Ir}$ the calibrated confidence values obtained by applying Platt scaling, temperature scaling, and isotonic regression, respectively.

The error-estimation methods are based on computing ℓ_1, ℓ_2, and ℓ_∞ norms of the prediction errors over confidence bins. Throughout this section, $\|\cdot\|_l$ denotes the ℓ_l norm, with $l \in \{1, 2, \infty\}$.

Let $\mathcal{D} = \{(s_i, y_i)\}_{i=1}^{N}$ denote a validation dataset, where $y_i \in \{0, 1\}$ indicates whether the i-th prediction is correct, and $s_i^c \in [0, 1]$ denotes the calibrated confidence associated with sample i obtained using calibrator $c \in \mathcal{C} \cup \{Uncalibrated\}$. For a fixed calibrator c, the interval $[0, 1]$ is partitioned into a set of disjoint confidence bins. Let $B \subseteq \{1, \ldots, N\}$ denote a confidence bin, i.e., a subset of indices corresponding to samples whose calibrated confidences s_i^c fall within a given interval, and let $|B|$ be its cardinality.

To capture deviation of confidence prediction errors within each bin using an ℓ_1-based metric, the Mean Absolute Error (MAE) quantifies the average deviation between calibrated confidence and empirical correctness within each bin. The bin-wise MAE is defined as:

$$\mathrm{MAE}_B(c) = \frac{1}{|B|} \sum_{i \in B} \|s_i^c - y_i\|_1,$$

To capture dispersion of confidence prediction errors within each bin using an ℓ_2-based metric, we compute the root mean squared error of the calibrated confidence values. The bin-wise Root Mean Squared Error (RMSE) is defined as:

$$\mathrm{RMSE}_B(c) = \frac{\left\| \left(s_i^c - y_i\right)_{i \in B} \right\|_2}{\sqrt{|B|}}$$

It is worth noting that, for binary outcomes, this dispersion metric corresponds to the square root of the bin-wise Brier score (BS_B): $\mathrm{RMSE}_B(c) = \sqrt{\mathrm{BS}_B(c)}$.

To capture dispersion of confidence prediction errors within each bin using an ℓ_∞-based metric, we compute the Maximum Absolute Error (MaxAE). This metric acts as a baseline indicator of worst-case behavior. For a given confidence bin B, the MaxAE is defined as:

$$\mathrm{MaxAE}_B(c) = \left\| \left(s_i^c - y_i\right)_{i \in B} \right\|_\infty$$

This metric captures the largest deviation between predicted confidence and observed correctness within a bin, and provides an ℓ_∞-based measure of worst-case uncertainty. While MAE and RMSE summarize average behavior, MaxAE highlights the most severe error and is particularly relevant for decision-making under guarantees.

As the optimal number of confidence bins that yield the most informative opinions is not known *a priori*, the above metrics are computed over a set of equal-width binning schemes, denoted by J; e.g., J ={5-bins,10-bins} when the range of values is partitioned into 5 and 10 intervals. Each binning scheme $j \in J$ induces a corresponding set of bins $B^{(j)}$, over which the metrics are evaluated.

Accordingly, the set of error estimations contains the three error estimation methods applied to each of the binning schemes:

$$Err = \bigcup_{j \in J} \{\text{MAE}_{B^{(j)}}, \text{RMSE}_{B^{(j)}}, \text{MaxAE}_{\text{B}^{(\text{j})}}\}$$

At present, the calibration techniques are induced over the calibration data. Updating the calibrators online is part of the future work.

4.2 Fuzzifier

The *Fuzzifier* is responsible for transforming the crisp confidence score produced by the Estimator, together with the set of opinions generated by the Estimator Performance module, into a fuzzy representation of confidence. Its purpose is to aggregate heterogeneous and possibly conflicting uncertainty information without collapsing it into a single scalar value. The benefit of using fuzzy theory for opinion aggregation is that the subsequent decision-making process can leverage its well-established mathematical foundations.

The output of the Fuzzifier is a set of $|\mathcal{C}| + 1$ fuzzy numbers that represent confidence together with its associated uncertainty. Unlike post-hoc calibration methods, which replace the original confidence score with a corrected scalar, the Fuzzifier preserves the original estimate and augments that information with an explicit representation of uncertainty based on the set of opinions on that confidence.

Construction of the Fuzzy Confidence: Our approach builds the fuzzy confidence using the α-cut method. This choice allows the aggregation of multiple opinions into a single, structured uncertainty representation while remaining compatible with standard fuzzy decision-making techniques.

Given the $\mathcal{C}$ and Err used in the concrete Estimator Performance, the input to the Fuzzifier consists of the collection of opinions *i* the crisp confidence scores s, s^{Pl}, s^{Ts}, and s^{Ir}, and *ii* the opinions derived from calibration and error-estimation methods, including calibrated confidence values and bin-wise error indicators such as MAE, RMSE, and MaxAE. Therefore, from each confidence estimate from the Estimator, the Fuzzifier receives as input ten opinions (three metrics computed across three binning schemes and the crisp confidence score) for each calibration method and the uncalibrated one. These opinions encode partial and complementary information about the reliability of s and may reflect different aspects of uncertainty.

Algorithm 1 describes the process for creating the fuzzy numbers from a collection of opinions. The procedure iterates over all calibrated variants of the

Algorithm 1. Create the fuzzy representation of the set of opinions

```
1: procedure FuzzifyOpinions
2:     fuzzyConfidences ← ∅
3:     for c ∈ C ∪ {Uncalibrated} do
4:         intervals ← ⟨⟩                                          ▷ EmptyList
5:         for j ∈ J do
6:             for e ∈ Err do
7:                 intervals ← intervals⊕createInterval(s^c, opinion(c, e, j))
                                                                   ▷ Append interval to a list
8:             end for
9:         end for
10:        intervals ← intervals⊕createSinglePointInterval(s^c)
                                                                   ▷ Append the crisp estimation
11:        sortedIntervals ← sort(intervals)
12:        cutIncrement ← 1/(length(intervals)-1)                  ▷ increment of each α-cut
13:        fuzzyConfidence ← EmptyFuzzySet
14:        α-cut ← 0
15:        for interval ∈ sortedIntervals do
16:            fuzzyConfidence ← addCut(fuzzyConfidence, α-cut, interval)
17:            α-cut ← α-cut + cutIncrement
18:        end for
19:        fuzzyConfidences ← fuzzyConfidences ∪ fuzzyConfidence
20:    end for
21:    return fuzzyConfidences
22: end procedure
```

confidence score, including the uncalibrated estimator output. Each of these iterations creates a fuzzy number in $[0, 1]$. For each calculated confidence s^c, the algorithm generates a set of intervals that represent the different opinions produced by the Estimator Performance module. These intervals are constructed by combining the calibrated confidence s^c with each error-based opinion across the considered binning schemes. Each opinion yields an interval $i \subseteq [0, 1]$ expressing the uncertainty associated with s^c under that specific view (line 7). In addition to the opinion-based intervals, the crisp confidence value s^c is added as an interval with zero width. This guarantees that the core of the resulting fuzzy confidence corresponds to the original estimate. Figure 2a exemplifies a list of generated intervals from an s^c. The full set of intervals is then sorted and used to construct an α-cut-based fuzzy number (lines 15-18). The fuzzy number that represents the fuzzy confidence is built incrementally by assigning increasing α-cut levels to the ordered intervals. The widest interval is associated with $\alpha = 0$, while successive intervals are assigned increasing α values with a uniform increment, ensuring that intervals that are contained in more opinions correspond to higher levels of membership. Figure 2b exemplifies *fuzzyConfidence* generated in that loop for the available opinions on an s^c. Finally, the algorithm returns the computed set of fuzzy numbers based on all received opinions.

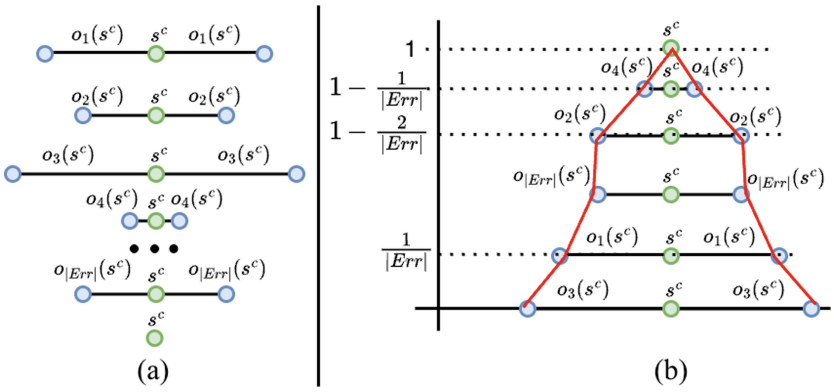


Fig. 2. Example of (a) list of intervals generated by the opinions on the confidence score and (b) fuzzy number that aggregates the intervals to represent a $fuzzyConfidence$.

4.3 Decision-Maker

The decision-maker leverages fuzzy confidence levels to reach a final decision. Figure 3a depicts an example of the four fuzzy confidences received by the decision-maker. While the development of advanced, problem-specific fuzzy decision models is beyond the scope of this work, we define here the mechanisms of the general decision-makers employed in the experimental section.

The decision-making process is structured into three distinct tasks. First, the system defines the inclusion criteria for the available fuzzy confidences. This involves determining whether to aggregate the entire set, exclude specific outliers (e.g., the most optimistic or pessimistic values), or select a single representative input (e.g., the most optimistic, pessimistic, or median).

Second, the decision-maker aggregates the selected information. Given that the inputs are fuzzy numbers, aggregation is performed either via fuzzy arithmetic operators (e.g., fuzzy sum and scalar division) or by averaging corresponding α-cuts. If the preceding selection phase filters out all but a single opinion, the aggregation function is the identity.

Third, to derive an actionable conclusion from the aggregated value, the decision-maker extracts real-valued descriptors that characterize the fuzzy number representative value and its associated uncertainty. Numerous real-valued descriptors have been proposed for defuzzification, ranking, and comparison [16,30]. Five traditional descriptors used in this study are: *i)* the midpoint of the *core* (m), *ii)* the *centroid* (e), *iii)* the *average width* of the α-cuts (w), *iv)* the area below a predefined threshold (z), and *v)* defuzzification based on the Kerre method (k).

These descriptors capture different facets of the fuzzy set. Descriptor m focuses exclusively on the central represented value and remains invariant to changes in uncertainty. Conversely, descriptor w serves as a pure measure of uncertainty and is invariant to horizontal translations of the fuzzy number.

Descriptors e, z, and k provide composite summaries, incorporating both the representative value and the degree of uncertainty. Figure 3b illustrates their values over a fuzzy number. Note that for the normal and convex fuzzy numbers used in this study, the average width of the α-cuts is mathematically equivalent to the total area under the membership function.

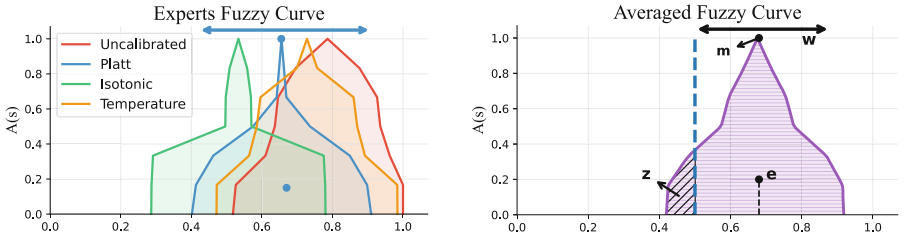


Fig. 3. Illustrative examples of (a) four fuzzy confidences received by the decision-maker (b) descriptor values of a fuzzy number.

5 Evaluation

This section demonstrates the feasibility of the UAFuCo principles and evaluates the quality of the pipeline decisions against the motivating scenario's needs.

5.1 Experiment Design

The experimental evaluation is designed to assess the quality of fuzzy confidence in autonomous decision-making across different goals, compared with decisions made using crisp confidence scores.

At the Estimator level, we employ a binary predictor trained on the drone flight trajectory dataset in [13] to classify paths as *safe* or *unsafe*. Each trajectory consists of 25 position samples and corresponding flight orientation parameters (x, y, z, r), which are subsequently subsampled to 12 points. We experiment with estimators implemented using a Support Vector Machine (SVM) with both *RBF* and *linear* kernels, with the regularization parameter set to $C \in \{1, 10\}$. We note that the scope of the experimentation is not to achieve the best possible estimations, but to enhance imperfect estimations and confidence scores with information about their uncertainty.

The calibrators in $\mathcal{C}$ used the set J ={3-bins, 5-bins, 10-bins} of binning schemes. Therefore, $|Err| = 3 \cdot 3 = 9$ and the Estimator Performance produces $(|\mathcal{C}| + 1) \cdot (|Err| + 1) = 40$ error-based opinions for each estimation.

The dataset is split into three disjoint subsets: training, calibration, and test. The training set is used to fit the Estimator model parameters; the calibration set

is used to train post-hoc calibrators in $\mathcal{C}$ and to compute error-based opinions; and the test set is reserved exclusively for final evaluation of decision-making performance.

To evaluate the decision-maker's performance within the context of the proposed rescue UAV scenario, the different operational priorities between mission phases must be considered. In the search phase (P_1), the system must prioritize locating the victim, which necessitates a *high-recall strategy* that emphasizes exploratory coverage and efficiency over extreme risk aversion. Conversely, as the mission transitions to the assistance phase (P_2), the requirements become strictly *risk-averse*, demanding a *high-precision approach* where navigation errors carry high costs.

To rigorously evaluate these evolving goals in our experiments, we utilize the F_β score as a primary metric. By varying the β parameter, we can quantitatively assess the system's effectiveness across the spectrum of mission requirements: using $\beta > 1$ to emphasize recall for the exploratory phase (P_1) and $\beta < 1$ to prioritize precision for the risk-averse phase (P_2). We experiment with $\frac{1}{9} \leq \beta \leq 9$. We also include the F_1 score for completeness, even though it does not favor any phase in the scenario.

The impact of fuzzy confidence on autonomous decision-making is evaluated by comparing it with the best possible decision-making that we identified for point confidence estimates. All decision-making experiments use the same candidate alternatives and predictions, isolating the effect of confidence representation from other sources of variation.

5.2 Performance of Crisp Decision-Makers

To conduct a comprehensive evaluation, we instantiated a separate crisp decision-maker for every combination of input type within $\mathcal{C} \cup \text{Uncalibrated}$, for each targeted F_β score, and for each SVM model we experimented with. This systematic approach resulted in a total of $4 \cdot 17 \cdot 4 = 272$ distinct decision-makers, allowing for a granular analysis of how different fuzzy representations align with specific operational priorities.

For each case, we evaluated the corresponding F_β performance under different thresholds for the confidence score in the decision-maker to decide whether a trajectory was $safe$ or $unsafe$ using the confidence score. Table 1 presents the results for the 68 decision-makers when the original estimation was produced by an SVM model with the linear kernel and $C = 1$. The table presents the threshold values for each crisp decision-maker; each threshold was identified by optimizing the corresponding F_β score. The results reveal a clear correlation between the β parameter and the optimal decision thresholds across all calibration methods. We observe that for high β values, where recall is prioritized, the optimal thresholds for deciding that a trajectory is $safe$ are significantly lower. This behavior aligns with the mission requirements of P_1, as a lower threshold reduces the decision-maker's strictness, ensuring that fewer potential trajectories are missed.

Conversely, as the β value decreases and the decision-maker shifts to favor precision, the optimal thresholds increase substantially to prioritize risk aversion.

Table 1. Best F_β performance by crisp decision-makers and the decision thresholds that achieved such a best score.

	β value																
	9	8	7	6	5	4	3	2	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$	$\frac{1}{9}$
Uncalibrated estimator																	
Best threshold	.04	.04	.04	.04	.04	.06	.06	.20	.39	.73	.73	.82	.84	.87	.87	.87	.87
Temperature scaling																	
Best threshold	.05	.05	.15	.15	.15	.20	.20	.32	.32	.38	.39	.39	.65	.69	.69	.69	.69
Platt scaling																	
Best threshold	.11	.11	.11	.12	.12	.22	.22	.22	.27	.27	.28	.28	.63	.63	.63	.63	.63
Isotonic regression																	
Best threshold	.02	.02	.02	.02	.03	.06	.06	.06	.27	.35	.35	.35	.56	.56	.56	.56	.56
F_β score	.97	.96	.96	.95	.94	.89	.83	.89	.72	.62	.61	.53	.55	.56	.56	.57	.57

In these scenarios, corresponding to the risk-averse phase (P_2), the maximum optimal thresholds range from 0.56 to 0.87. This high-threshold strategy minimizes the risk of false positives.

The uncalibrated estimator exhibits the widest variance in optimal thresholds, spanning nearly the entire unit interval, whereas the calibrated methods demonstrate more constrained and stable threshold shifts across the spectrum of mission requirements. Although each post-calibration action updates the resulting estimation confidence, with the expectation of bringing it closer to its true value, the decision-maker F_β is extremely similar for all four methods because the optimal threshold identified for each F_β has also been adjusted accordingly. The last row in Table 1 shows the shared F_β scores.

5.3 Design Details of the Fuzzy Decision-Maker

As the primary objective of this work is not to optimize fuzzy decision-making architectures but to demonstrate that improvements over crisp baselines are readily attainable, the logic implemented within the fuzzy decision-makers remains intentionally straightforward.

The inclusion criteria for the first task are based on extremum functions. For decisions within mission phase P_1, the system selects the most optimistic fuzzy confidence regarding trajectory safety; conversely, for P_2, the most pessimistic value is chosen. Consequently, the aggregation phase for these selected confidences reduces to the identity function. The third task extracts the five descriptors introduced in Sect. 4.3 and maps them to a latent confidence measure in

$[0, 1]$. This mapping is performed via a logistic regression model trained on the five descriptors and the ground-truth labels. The optimal decision threshold is subsequently determined by maximizing the corresponding F_β score.

Empirical results indicate that the fuzzy decision-maker outperforms crisp alternatives for all "extreme" β values. However, performance decreases in the intermediate regime, specifically for $\frac{1}{3} \leq \beta \leq 3$. The results suggest that utilizing a radical selection strategy (i.e., strictly the most pessimistic or optimistic values) is inappropriate when the balance between precision and recall is not heavily skewed. To address this, for the range $\frac{1}{3} \leq \beta \leq 3$, we evaluated a configuration that retains all fuzzy confidences in the first task and aggregates them via an averaging operator in the second task.

Table 2. F_β for the fuzzy decision-maker, and the best threshold found for the latent confidence measure

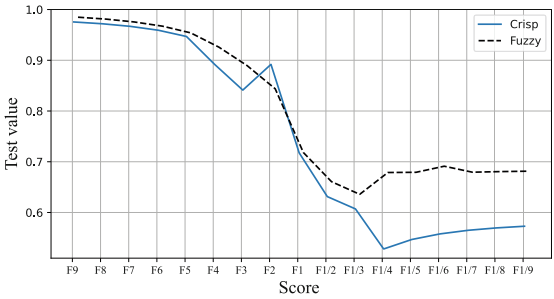
	β value																
	9	8	7	6	5	4	3	2	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$	$\frac{1}{9}$
Fuzzy decision-maker																	
Threshold	.04	.04	.04	.04	.04	.05	.21	.21	.33	.39	.56	.59	.59	.59	.59	.59	.59
F_β score	.99	.98	.98	.97	.95	.93	.89	.84	.72	.66	.64	.68	.68	.69	.68	.68	.68

5.4 Evaluation of the Fuzzy Decision-Maker

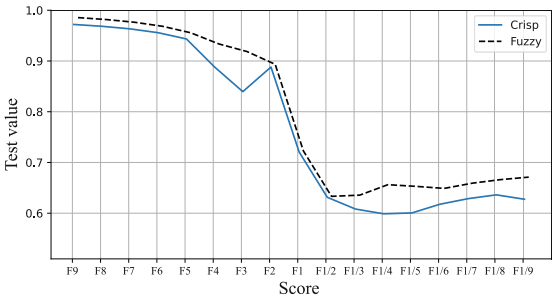
Table 2 summarizes the F_β scores for the fuzzy decision-maker, utilizing an SVM model with a linear kernel ($C = 1$) for the initial estimation. The results indicate that the fuzzy decision-maker outperforms its crisp counterparts across nearly all β values, with the exceptions of F_1, where performance was identical, and F_2. We also report the optimal thresholds for the latent confidence derived from the fuzzy descriptors; as hypothesized, these thresholds exhibit a monotonic increase as β decreases, reflecting a shift toward higher precision requirements.

Figure 4a provides a visual comparison of the results from Tables 1 and 2, highlighting the performance gaps between the two paradigms. A significant improvement is observed for $\beta = 3$, a trend consistent across several cases. For higher β values, the gap narrows to approximately zero as both F-scores increase and approach one. Conversely, in the risk-averse phase ($\beta < 1$), the proposed approach yields greater gains. For instance, at $\beta = \frac{1}{4}$, the fuzzy decision-maker achieves a gap of 0.15.

Experiments conducted with a linear kernel and $C = 10$ (Fig. 4b) display a similar trend. In this configuration, the fuzzy decision-maker outperforms the crisp baseline for all tested β values, although the average performance margin is narrower than in the $C = 1$ case, particularly in the risk-averse scenarios ($\beta < 1$).



(a) Linear kernel, $C = 1$.



(b) Linear kernel, $C = 10$.

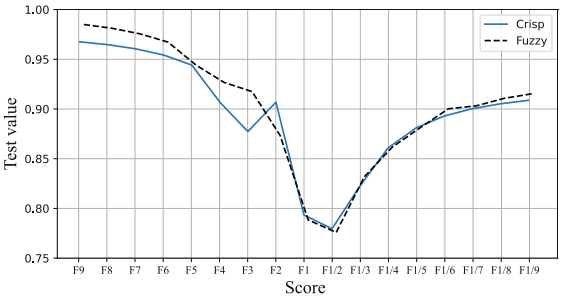
Fig. 4. F_β for fuzzy and crisp decision-makers with linear kernel SVM.

To determine whether these trends persist in different estimation models, we replicated the experiments using an RBF kernel. The results, depicted in Fig. 5, corroborate the previous observations. The fuzzy decision-maker remains superior for $\beta > 2$, while for $\beta < 1$, it consistently meets or exceeds the performance of the crisp baselines (with $C = 1$ and $C = 10$, respectively). Although the performance gains observed with the RBF kernel are more modest than those with the linear kernel, the gaps remain distinct. These results demonstrate that even a baseline fuzzy implementation can outperform optimized crisp decision-makers. Furthermore, the flexibility to tailor uncertainty descriptors and the parameters that combine them for specific systems suggests significant potential for future research into specialized fuzzy decision-makers.

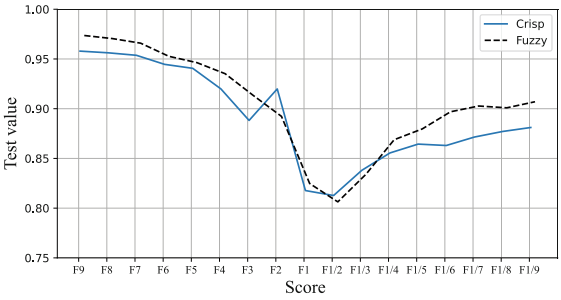
The main takeaway from our observations is that the fuzzy decision maker outperforms the best crisp decision maker in most cases, with the performance gap widening as the preference for either precision or recall becomes more pronounced.

5.5 Interpretability of Decisions

We further evaluate the interpretability of UAFuCo by analyzing instances of prediction failure. In cases of misclassification, the visual representation of fuzzy



(a) RBF kernel, $C = 1$.



(b) RBF kernel, $C = 10$.

Fig. 5. F_β for fuzzy and crisp decision-makers with RBF kernel SVM.

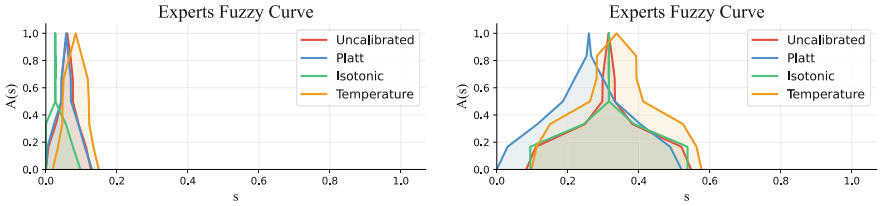
numbers clearly shows whether the estimator output lacks sufficient evidence to support the correct decision, even when accounting for typical estimation error and uncertainty. Two illustrative cases are detailed below.

Figure 6a illustrates a scenario where all fuzzy opinions strongly support an erroneous prediction. This is evidenced by cores positioned close to zero and low uncertainty, indicated by the narrow spread of the fuzzy numbers. In this instance, the uncalibrated crisp score was 0.06, which provides some insight into the error; however, the fuzzy representation confirms that the model was highly confident in its incorrect assessment.

Conversely, Fig. 6b depicts a case with an uncalibrated crisp score of 0.315. As this score approaches values that might be correctly classified under different thresholds, it is essential to examine whether any evidence supported the correct class given the potential estimation errors. The fuzzy representation reveals that despite higher uncertainty (indicated by wider fuzzy numbers), the vast majority of the supports are distributed within the $[0, 0.5]$ interval. This indicates that none of the expert perspectives found the true class plausible, providing a transparent justification for the failure.

Such granular information regarding the uncertainty surrounding a point estimate is inevitably lost when using crisp values. UAFuCo not only generates a final prediction but also exposes the underlying uncertainty of incorrect decisions

through the visual representation of fuzzy numbers, rendering failure cases more transparent and interpretable to the end user.



(a) Low uncertainty misclassification (b) High uncertainty misclassification

Fig. 6. Samples for the SVM model that failed the decision.

6 Conclusion

This paper addressed the critical challenge of miscalibrated confidence scores in autonomous systems by proposing UAFuCo, an approach that is"uncertainty-aware by construction." While post-hoc calibration methods generally improve the reliability of point estimates, a single scalar value inevitably hides the complex uncertainty involved in calculating the value. Therefore, they are insufficient for robust decision-making under prediction uncertainty and changing operational goals.

UAFuCo preserves heterogeneous reliability indicators for expected errors in computing the confidence score by aggregating them and presenting them as fuzzy numbers. This representation effectively encodes second-order uncertainty, allowing the system to maintain a transparent and mathematically grounded view of its own predictive reliability. The initial experimental results with simple decision-makers demonstrate that this fuzzy representation provides an advantage in goal-dependent decision-making over optimal decision-makers that reason with a single scalar value. Furthermore, the proposed architecture is estimator-agnostic, serving as a non-intrusive wrapper that enhances the interpretability and adaptability of existing ML components without requiring access to their internal structures. Our findings suggest that fuzzy theory offers a practical middle ground between oversimplified scalar values and ad hoc methods for communicating uncertainty to the decision-maker.

As future work, we plan to conduct ablation studies to analyze the contribution of each opinion type to the quality and expressiveness of fuzzy confidence and to explore online calibration updates. We also plan to investigate optimal decision-makers using the fuzzy inputs of the proposed approach.

Acknowledgments. This work was partially supported by the Helmholtz Association with the pilot program Core Informatics at KIT (KiKIT) and Helmholtz Grant 46.23

(Engineering Secure Systems) and by the German Research Foundation (DFG) - SFB 1608 - 501798263.

References

1. Braga, P.L., Oliveira, A.L.I., Meira, S.R.L.: Software effort estimation using machine learning techniques with robust confidence intervals. In: 7th International Conference on Hybrid Intelligent Systems (HIS 2007), pp. 352–357 (2007)
2. Chen, S., Deng, Y., J.W.: Fuzzy sensor fusion based on evidence theory and its application. Appl. Artif. Intell. **27**(3), 235–248 (2013)
3. d'Avila Garcez, A.S., Gori, M., Lamb, L., Serafini, L., Spranger, M., Tran, S.: Neural-symbolic computing: an effective methodology for principled integration of machine learning and reasoning. FLAP **6**, 611–632 (2019)
4. Del Corso, G., Colantonio, S., Caudai, C.: Shedding light on uncertainties in machine learning: formal derivation and optimal model selection. J. Franklin Inst. **362**(3), 107548 (2025)
5. Ding, Z., Han, X., Liu, P., Niethammer, M.: Local temperature scaling for probability calibration. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6869–6879 (2021)
6. Fung, M.L., Chen, M.Z.Q., Chen, Y.H.: Sensor fusion: a review of methods and applications. In: 2017 29th Chinese Control And Decision Conference (CCDC), pp. 3853–3860 (2017)
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning, ICML 2016, vol. 48, pp. 1050–1059. JMLR.org (2016)
8. Ghoushchi, S.J., Mardani, A., Martínez, L.: Trust number: trust-based modeling for handling decision-making problems. Knowl.-Based Syst. **305**, 112631 (2024)
9. Grassi, V., Mirandola, R., Perez-Palacin, D.: Embracing uncertainty: a fuzzy theoretical model for goal fulfillment assessment. In: 2025 IEEE/ACM 13th International Conference on Formal Methods in Software Engineering (FormaliSE), pp. 113–123 (2025)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, vol. 70, pp. 1321–1330. JMLR.org (2017)
11. Kamis, N.H., Abdullah, K., Mohamed, H., Sudin, S., Ishak, W.Z.A.W.: Decision making models based on consistent fuzzy preference relations with different defuzzification methods. In: 2011 IEEE Colloquium on Humanities, Science and Engineering, pp. 845–850 (2011)
12. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 5580–5590. Curran Associates Inc, Red Hook, NY, USA (2017)
13. Khatiri, S., Mohammadi Amin, F., Panichella, S., Tonella, P.: When uncertainty leads to unsafety: empirical insights into the role of uncertainty in unmanned aerial vehicle safety. Empirical Softw. Eng. **30**(6), 166 (2025)
14. Klir, G.J., Yuan, B.: Fuzzy sets and fuzzy logic: theory and applications. Prentice Hall, Upper Saddle River, NJ (1995)

15. Mauris, G., Lasserre, V., Foulloy, L.: A fuzzy approach for the expression of uncertainty in measurement. Measurement **29**(3), 165–177 (2001)
16. McCahon, C.S., Stanley Lee, E.: Comparing fuzzy numbers: the proportion of the optimum method. Inter. J. Approximate Reasoning **4**(3), 159–181 (1990)
17. Mohamad, D., Sulaiman, N.H., Abu Bakar, A.S.: Solving industrial decision making problems using fuzzy numbers. In: 2011 IEEE Symposium on Business, Engineering and Industrial Applications (ISBEIA), pp. 271–274 (2011)
18. Alonso Moral, J.M., Castiello, C., Magdalena, L., Mencar, C.: Explainable Fuzzy Systems. SCI, vol. 970. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71098-9
19. Neal, R.M.: Bayesian Learning for Neural Networks. Springer-Verlag, Berlin, Heidelberg (1996)
20. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, pp. 625–632. ACM (2005)
21. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press (1999)
22. Rousseau, A.-J., Becker, T., Appeltans, S., Blaschko, M., Valkenborg, D.: Post hoc calibration of medical segmentation models. Dis. Appli. Sci. **7**(3), 180 (2025)
23. Sasiadek, J.: Sensor fusion. Annu. Rev. Control. **26**(2), 203–228 (2002)
24. Sbeyti, M.K., Karg, M., Wirth, C., Klein, N., Albayrak, S.: Cost-sensitive uncertainty-based failure recognition for object detection. In: Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence. JMLR.org (2024)
25. Tyralis, H., Papacharalampous, G.: A review of predictive uncertainty estimation with machine learning. Artif. Intell. Rev. **57**(4), 94 (2024)
26. Weyns, D., Calinescu, R., Mirandola, R., Tei, K., et al.: Towards a research agenda for understanding and managing uncertainty in self-adaptive systems. SIGSOFT Softw. Eng. Notes **48**(4), 20–36 (Oct.2023)
27. Yoshida, Y., Yasuda, M., Ichi Nakagami, J., Kurano, M.: A new evaluation of mean value for fuzzy numbers and its application to American put option under uncertainty. Fuzzy Sets Syst. **157**(19), 2614–2626 (2006), Fuzzy Sets and Probability/Statistics Theories
28. Zadeh, L.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
29. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 694–699. Association for Computing Machinery, New York, USA (2002)
30. Zhu, Q., Lee, E.: Comparison and ranking of fuzzy numbers. In: Fuzzy Regression Analysis, vol. 1, pp. 21–44. Omnitech Press (1992)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.