Uncertainty Discourse and Applications

www.uda.reapress.com

Uncert. Disc. Appl. Vol. 3, No. 4 (2026) 320–332.

Paper Type: Original Article

# Evidence Weighted Fuzzy Fusion with Supervised Reliability Estimation under Controlled Source Corruption

**Reza Hadavi**[1], **Setayesh Zahiri**[2], **Mahmoud Alimoradi**[1,*]

[1] Department of Computer Engineering, Ayandegan University, Tonekabon, Iran; reza.hadavi@gmail.com; mahmoud.alimoradi@aihe.ac.ir.

[2] Morvarid Intelligent Industrial Systems Research Group, Iran; setayeshzahiri70@gmail.com.

**Citation:**

| | |
|---|---|
| *Received: 20 April 2026*<br>*Revised: 13 July 2026*<br>*Accepted: 10 September 2026* | Hadavi, R., Zahiri, S., & Alimoradi, M. (2026). Evidence weighted fuzzy fusion with supervised reliability estimation under controlled source corruption. *Uncertainty Discourse and Applications*, *3*(4), 320-332. |

## Abstract

This study compares evidence-weighted fusion of classifier outputs under controlled corruption of one source. Essential Work of Fracture (EWF)-Fuse uses reliability weights estimated from labelled validation accuracies, discounted evidence combination, consensus-based batch adaptation, and a conflict-triggered averaging fallback. Three public tabular datasets are partitioned into three feature blocks with distinct classifiers and evaluated over ten seeds per dataset and corruption condition. At 30% source corruption, the reported pooled mean accuracy is 90.2% for EWF-Fuse, 77.6% for raw Dempster–Shafer combination, 84.7% for probability averaging, 82.9% for majority voting, and 93.3% for stacking. Removing learned weights reduces EWF-Fuse accuracy to 78.0%. The reported paired tests favor EWF-Fuse over the three classical fusion comparators but require caution because repeated splits overlap within only three datasets. At 20% corruption, EWF-Fuse has the highest Expected Calibration Error (ECE) (0.319) among the compared methods. A separate conceptual specification describes consonant mapping, agreement-based unsupervised weighting, and provenance records; the reported experiments do not validate those components. The findings support a supervised proof of concept for reliability weighting and identify calibration, implementation provenance, and dependence-aware evaluation as priorities.

**Keywords:** Multisource data fusion, Dempster-Shafer evidence theory, Fuzzy membership functions, Source reliability estimation, Conflict management, Reproducible data science.

## 1 | Introduction

A monitoring pipeline may combine assertions from sensors, databases, and human reports that disagree about the same entity. Such disagreement can arise from measurement error, model bias, different observation windows, or incompatible meanings. Reliable fusion therefore requires both a shared interpretation of the candidate states and an explicit account of source quality. The present experiments instantiate this problem

Corresponding Author: mahmoud.alimoradi@aihe.ac.ir

https://doi.org/10.48313/uda.v3i4.118

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

through heterogeneous classifiers on feature blocks; they do not test air-pollution advisories or other safety-critical decisions.

Combining uncertain assertions from multiple sensors and databases has been studied for decades, and multisensor fusion is a mature engineering discipline [1]. Fuzzy set theory contributed graded membership, which represents vague or linguistically expressed assertions more faithfully than binary predicates [2]. Dempster-Shafer evidence theory represents support as belief masses over sets of hypotheses and provides a calculus for combining them [3], [4]. In parallel, the truth-discovery literature studies how to identify correct values among conflicting claims from sources of unknown trustworthiness [5]. What remains difficult is integrating these threads into one pipeline that is uncertainty-aware, reliability-aware, and auditable by third parties.

Dempster's rule can become sensitive to conflict normalization when highly conflicting evidence leaves little common support, as illustrated by Zadeh's example [6]. Yager's rule reallocates conflict to ignorance [7], while Murphy's rule averages masses before repeated combination [8]. The transferable belief model separates belief representation from decision probabilities [9]. Reliability discounting provides an additional way to reduce a source's evidential influence [10]. These mechanisms motivate the present comparison, but no combination rule resolves incorrect shared semantics or all forms of source dependence.

The manuscript distinguishes a conceptual fusion architecture from the supervised implementation reported in the experimental campaign. The architecture specifies a normalized possibility-to-consonant mapping, agreement-based weighting, and provenance records. The evaluated implementation instead obtains weights from labelled validation accuracies and describes entropy-strength evidence assignments with direct discounted combination and a fallback. Keeping these layers explicit identifies which claims the reported outcomes can support.

The conceptual formulation provides a consonant representation of normalized possibility profiles, states the reliability-discounting and decision equations, and identifies provenance information needed for replay. The experimental contribution is a supervised comparison on three feature-partitioned datasets under paired feature dropout and training-label corruption. Ablations examine weight sharpness, consensus adaptation, and fallback use; the reported results show that static reliability weights account for most of the observed accuracy gain.

Section 2 reviews the relevant representation and combination literature. Sections 3 and 4 describe the conceptual specification and its scope. Section 5 separately states the reported supervised experimental realization and outcomes. Section 6 identifies limitations, and Section 7 draws conclusions at the scale supported by the campaign.

# 2 | Related Work

## 2.1 | Uncertainty Formalisms for Graded Assertions

Fuzzy memberships represent graded predicates [2]. When such grades are interpreted as a normalized possibility distribution, they can be represented by consonant belief functions [11]. This semantic assumption is substantive: class probabilities and possibility grades are distinct objects, and simply rescaling a probability vector to maximum one does not preserve its probabilistic interpretation.

The representation must be chosen for the meaning of the input assertions. In the conceptual specification, normalized possibility grades describe singleton plausibilities; in the reported experiments, classifier probability vectors supply numerical inputs. The experiments therefore examine the resulting implemented fusion scores rather than validate a general conversion from linguistic vagueness into belief masses. The Awareness Logic of Ambiguity distinguishes interpretive valuation, contextual alignment, and epistemic restraint [12]; this supplies a semantic reason to record assertion meaning separately from learned source reliability, without replacing the evaluated fusion rule.

Dempster's multivalued mappings induce lower and upper probabilities [3], and Shafer's theory provides belief assignments on a frame of discernment [4]. The transferable belief model distinguishes credal representation from pignistic decision making [9]. We use that distinction to state mass-level discounting separately from the final read-out; it does not establish probability calibration.

## 2.2 | Combination Rules and Conflict Management

Dempster's rule combines independent evidence by multiplying the masses of intersecting focal sets and renormalizing by one minus the conflict coefficient. The rule is commutative, associative, and justified when its assumptions hold, which explains its canonical status [3], [4]. Its normalization step, however, redistributes conflict proportionally among the remaining focal sets, and Zadeh showed that slight overlap between two highly conflicting but reliable sources can be inflated into certainty [6]. This pathology motivated a family of alternative rules that differ in where the conflict mass is deposited [7], [11].

Yager and Dubois–Prade rules differ in how they reallocate conflict [7], [11]. Murphy's repeated-combination construction first averages masses [8]. Denoeux's cautious combination concerns nondistinct bodies of evidence [13] and is a different operator. Averaging can moderate outlier evidence, but repeated Dempster self-combination still sharpens the pooled mass and can magnify systematic errors.

Contextual discounting models reliability that varies with the state of the world [10]. The present source-wise weights are a simpler scalar choice. Our comparison examines their effect within one implemented fusion pipeline; it does not establish that weighted pre-averaging is universally preferable to other conflict-management rules.

## 2.3 | Source Reliability, Truth Discovery, and Reproducible Pipelines

Truth-discovery research studies conflicting claims from sources of unknown quality [5], [14]. Agreement-based estimators connect inferred source quality to inferred truth, but copying and other dependencies can distort agreement [15]. These ideas motivate the conceptual unsupervised component. The reported experiments use supervised validation accuracy instead and do not evaluate this fixed-point estimator.

A provenance record should retain the inputs, transformations, weighting parameters, and decision rule needed to explain a fused output. This design treats the estimated weights and their versions as part of the record. Recording them supports inspection; it does not make those weights identifiable as true source accuracies.

The Open Provenance Model supplies a vocabulary for artifacts, processes, and agents [16], while reproducible computational research requires inspectable inputs and executable procedures [17]. The present architecture adopts these as design requirements. The experimental description reports a rerun check and limited logs, rather than a complete independently audited replay system.

## 2.4 | Synthesis and Positioning

Across the reviewed streams, four requirements emerge for trustworthy fusion of conflicting assertions. The representation layer must express graded assertions without pretending to more precision than the sources possess, and the weighting layer must estimate reliability without ground truth while degrading gracefully when its assumptions fail. The combination layer must resist conflict pathologies while remaining interpretable. The systems layer must emit decisions that are versioned, deterministic, and auditable by third parties.

The framework connects representation, weighting, combination, and provenance as an engineering specification. The contribution is their explicit separation and a supervised proof-of-concept comparison, rather than a verified claim that no prior framework meets these requirements.

# 3 | Problem Formulation and Research Questions

## 3.1 | Research Questions

Three questions organize the analysis. RQ1 concerns the mathematical and semantic conditions for a consonant mapping. RQ2 concerns how source-quality weights influence fusion under heterogeneous corruption. RQ3 concerns what inputs and records are needed to reproduce a decision. The reported experiments address a supervised realization of RQ2 and provide only partial evidence on RQ3.

The empirical comparison examines whether reliability-weighted fusion loses less accuracy than raw Dempster combination as one source is corrupted, and whether learned weights improve on uniform weights. It also reports a deterministic rerun check. The unsupervised fixed-point estimator, its convergence, and a complete provenance replay were not evaluated, so the campaign does not validate a universal unsupervised reliability guarantee or necessary-and-sufficient audit conditions.

## 3.2 | System Model and Notation

For each entity e, let Θ_e be a finite, exhaustive frame of mutually exclusive candidates. Sources s ∈ S supply profiles on that same frame. In the conceptual specification these are possibility grades; α_s ∈ [0, 1] is a reliability-discounting coefficient, β_s is a normalized averaging weight, m_s a mass assignment, and G a provenance ledger. The supervised experimental realization is specified separately in Section 5.

## 3.3 | Formal Problem Statement

Order a nonnegative profile as μ_(1) ≥ … ≥ μ_(K) and normalize its positive maximum to one. Let B_(k) contain the k highest-ranked candidates, with a deterministic tie order. Interpreting the normalized grades as singleton plausibilities gives the consonant assignment [11].

$$\mathrm{m_s(B_(k))} = \mathrm{mu_(k)} - \mathrm{mu_(k+1)}, \mathrm{k} = 1, \ldots, \mathrm{K}, \text{with } \mathrm{mu_(K+1)} = 0,\ \mathrm{B_(K)} = \mathrm{Theta_e}. \quad \textbf{(1)}$$

*Eq. (1)* telescopes to total mass one; tied adjacent grades generate zero masses, so the representation is unaffected by their ordering after zero-mass focal sets are removed. Reliability discounting uses α_s ∈ [0, 1], moving the undiscounted complement to Θ_e [4], [10]. *Eq. (3)* defines pairwise conflict.

$$\mathrm{m_s^{alpha_s}(A)} = \mathrm{alpha_s} * \mathrm{m_s(A)} \text{ for } \mathrm{A} != \mathrm{Theta_e};\ \ \mathrm{m_s^{alpha_s}(Theta_e)} = 1 - \mathrm{alpha_s} + \mathrm{alpha_s} * \mathrm{m_s(Theta_e)}. \quad \textbf{(2)}$$

$$\mathrm{K_\{ij\}} = \mathrm{sum_\{A\ intersect\ B\ =\ empty\ set\}\ m_\{s_i\}^\{alpha_i\}(A)} * \mathrm{m_\{s_j\}^\{alpha_j\}(B)}. \quad \textbf{(3)}$$

$$\tilde{\mathrm{m}} = \Sigma_\mathrm{s}\ \beta_\mathrm{s}\ \mathrm{m_s^\{\alpha_s\}}, \beta_\mathrm{s} = \alpha_\mathrm{s} / \Sigma_\mathrm{j}\ \alpha_\mathrm{j} \text{ when } \Sigma_\mathrm{j}\ \alpha_\mathrm{j} > 0;\ \mathrm{m_fuse} = \tilde{\mathrm{m}} \oplus \cdots \oplus \tilde{\mathrm{m}} \text{ (n factors). If } \Sigma_\mathrm{j}\ \alpha_\mathrm{j} = 0, \text{set } \mathrm{m_fuse(\Theta_e)} = 1. \quad \textbf{(4)}$$

In *Eq. (4)*, ⊕ denotes normalized Dempster combination, defined only when its conflict is less than one. Repeating the average is a Murphy-style sharpening operation [8], not independent acquisition of new evidence and not cautious combination. If a normalization is undefined, the specification returns the vacuous mass on Θ_e and records that fallback. For normalized masses with m_fuse(∅) = 0, the pignistic transformation distributes each focal mass over its elements [9].

$$\mathrm{BetP(theta_k)} = \mathrm{sum_\{A\ contains\ theta_k\}\ m_fuse(A)} / |\mathrm{A}|. \quad \textbf{(5)}$$

The conceptual task is to choose a reliability-estimation procedure and fusion rule for aligned profiles while recording sufficient inputs and transformations for replay. Without labels or an identifiable generative model,

expected decision error is unknown and is not a jointly solved optimization objective. *Eq. (4)* defines one candidate rule; Section 5 reports a different direct-combination realization.

# 4 | Proposed Methodology

## 4.1 | Framework Overview

The conceptual architecture separates ingestion, representation, reliability estimation, combination, and decision recording. This separation identifies interface requirements and allows components to be evaluated independently. It is a proposed specification rather than evidence that a full production pipeline was implemented.

Ingestion projects assertions onto an explicitly defined common frame and records units, context, and rejected inputs. Mass construction then applies *Eq. (1)* only when a possibility interpretation is justified. Agreement-based reliability estimation is a conceptual option; the experiment uses validation-based weighting.

The combination stage records discounting coefficients and conflict values, and the decision stage records the pignistic output and tie rule. These records define intended replay inputs. The limited experimental logs described in Section 5 exercise only part of this architecture.

## 4.2 | From Fuzzy Assertions to Consonant Belief Structures

*Eq. (1)* requires finite nonnegative grades, a positive maximum, and an explicitly adopted possibility interpretation. A positive unnormalized profile can be rescaled, with the factor recorded; an all-zero profile cannot and is rejected or represented as ignorance under a documented convention. Ranking any such profile produces nested focal sets, including profiles with separated positive peaks. Existing non-consonant mass assignments are different inputs and cannot be recovered from singleton plausibilities alone. Different sources must share an aligned frame and compatible meanings; mathematical normalization does not establish semantic commensurability.

In this construction $\mu_k$ is the plausibility of singleton $\theta_k$, not its belief or probability [11]. The consonant mass reproduces the profile's singleton plausibilities and does not independently validate the source's judgments. Classifier probabilities used in the experiments require a separate mass-construction convention; their numerical values alone do not supply the possibility semantics assumed here.

## 4.3 | Agreement-Based Reliability Estimation

The conceptual agreement estimator compares pignistic profiles on overlapping entities and proposes updates that increase with weighted concordance [5], [14]. For independent draws from profiles $p_i$ and $p_j$, agreement is $\Sigma_k p_i(k)p_j(k)$, averaged over overlap. This statistic measures concordance, not accuracy. The manuscript supplies no complete normalized update equation or convergence proof, and the experimental campaign does not execute this estimator.

Dependence checks and leave-one-source-out sensitivity analyses could help diagnose copied or fragile evidence [15]. They do not identify a colluding but internally consistent false majority without an external anchor. The reported experiment does not validate such safeguards, so their update equations, thresholds, convergence, and failure handling remain implementation requirements.

## 4.4 | Conflict-Aware Weighted Combination

The conceptual rule discounts source masses, averages them with normalized $\beta_s$, and repeatedly combines the average. Averaging is idempotent as a pooling operation when identical masses are supplied, but the full repeated-combination rule is generally not. Self-combination can sharpen a pooled profile despite dependence, and its interpretation must distinguish this numerical transformation from fusion of independent observations.

Discounting reduces committed mass before combination, but high disagreement does not guarantee convergence toward maximal ignorance. For two fully reliable point masses on different singleton candidates, the equal average has mass 0.5 on each singleton and zero on the full frame; its two-fold normalized self-combination retains those masses. Thus it yields a non-vacuous tie rather than full ignorance. Source order is immaterial to the specified fixed-weight average, but it can matter if weights are adapted sequentially. The supervised experiment adds a different high-conflict fallback described in Section 5.

## 4.5 | Decision Layer and Provenance Ledger

A proposed provenance record contains input hashes, reliability estimates and traces, conflict values, the rule identifier, code and environment versions, random seeds, and the output [16], [17]. Replay also requires the original input bytes, ordering conventions, and numerical environment. Fixed ordering and seeds help reproducibility but do not alone guarantee bitwise identity across platforms. The reported experiment claims two identical result files; a complete independent audit of G is not supplied.

## 4.6 | Complexity, Assumptions, and Failure Modes

Pairwise concordance over M overlapping entities costs $O(n^2MK)$, and constructing each ranked consonant profile costs $O(K \log K)$. Each source has at most K nonzero focal sets. Averaging sources with different candidate orders and taking repeated intersections can destroy consonance and produce many more focal sets, up to an exponential dependence on frame size in general. Complexity therefore depends on the actual focal-set representation and pruning policy. Aligned frames, informative source diversity, and complete retained records are assumptions; coherent collusion or unlogged input errors need not be detectable from internal logs.

**Algorithm 1. Conceptual fusion specification with optional agreement weighting.**

Input: aligned profiles μ_{s,e}, finite frame Θ_e, supplied coefficients α_s ∈ [0,1], deterministic tie order
1: for each entity e do
2: validate nonnegative finite profiles and their common possibility interpretation
3: if any profile has maximum zero, record rejection and continue with the next entity; otherwise normalize profiles to maximum one
4: build each consonant mass m_s using *Eq. (1)*
5: record the estimator, data split and version associated with every supplied α_s
6: if Σ_s α_s = 0, set m_fuse(Θ_e) = 1 and go to step 12
7: discount each mass using *Eq. (2)*
8: set β_s = α_s / Σ_j α_j
9: set m = Σ_s β_s m_s^{α_s}
10: combine m with itself n times by normalized Dempster combination
11: if a normalization is undefined, replace the output by the vacuous mass and record the fallback
12: compute BetP using *Eq. (5)*
13: select its maximum using the declared deterministic tie order
14: record inputs, coefficients, conflicts, rule, versions and output in G
15: end entity loop
16: return decisions and records
Output: conceptual fused estimates and their provenance records

# 5 | Experimental Results and Analysis

## 5.1 | Experimental Setup

The reported experiment uses breast cancer (569 instances, 30 features, 2 classes), optical digits (1,797 instances, 64 features, 10 classes), and wine (178 instances, 13 features, 3 classes) from scikit-learn [18], [19]. Features are standardized with inner-training statistics and partitioned into contiguous blocks of widths 10/10/10, 22/21/21, and 5/4/4. Sources 1–3 use logistic regression ($C = 1$), seven-nearest-neighbor classification, and Gaussian naive Bayes. These are constructed classifier sources on a shared dataset, not independently acquired sensors or human assertions.

Corruption is injected at source 2 only, at four levels c in {0, 10, 20, 30}%, combining two fault modes. Feature dropout sets a c-fraction of source-2 entries to zero after standardization - that is, replaces them with the feature mean - and is applied to the inner-training, validation, and test partitions alike, emulating a sensor whose readings drop out persistently rather than only during model fitting. Label noise flips a c-fraction of source-2's training-fold labels to a different class, so the corrupted source's own model is also fitted on mislabelled data; the labels seen by the other sources, by the validation split, and by the test evaluation remain clean. The source reports seeded generators for the splits and corruption draws.

Each of ten seeds uses a stratified 75/25 train/test split and an 80/20 inner-training/validation split of the training portion. The reported Essential Work of Fracture (EWF)-Fuse realization learns static weights proportional to $acc_s^2$ from clean validation labels; it is supervised. It then adapts effective strengths through consensus-referenced exponential averaging (rate 0.3) over ten test batches and uses reliability-weighted averaging when instance-level raw pairwise conflict exceeds $\tau = 0.5$. The description also reports evidence strength $1-H(p)/\log K$ for source probability vectors, direct discounted Dempster combination, and pignistic read-out. It does not supply the exact focal-mass allocation or establish equivalence to the consonant mapping and repeated-average combination of *Eqs. (1)* and *(4)*. The empirical results therefore concern this described implementation rather than validate the conceptual algorithm. Averaging and voting are conventional comparators [20]; stacking fits a logistic meta-learner ($C = 1$) to concatenated validation probabilities [21]. Expected Calibration Error (ECE) uses ten equal-width bins and is reported alongside Negative Log-Likelihood (NLL) [22].

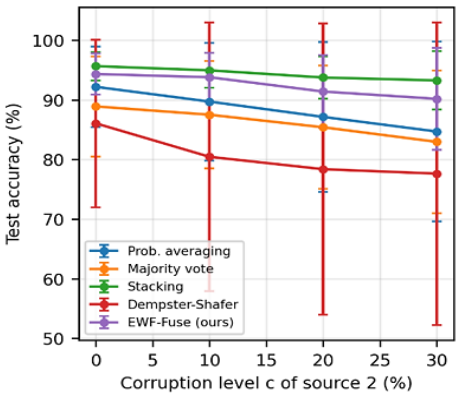
The source reports a 27-second campaign on three datasets, four corruption levels, ten seeds, five methods, and six ablation variants under Python 3.12 with NumPy, SciPy, scikit-learn, statsmodels, and Matplotlib. It also reports two byte-identical result files with SHA-256 prefix 07a96ea9c2a0. The executable files and complete hashes are not established by the supplied DOCX. These statements remain reported reproducibility observations rather than an independently rerun result or complete provenance audit.

## 5.2 | Main Comparison

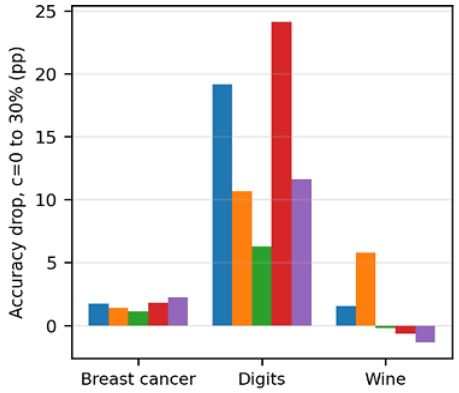
*Table 1* reports pooled accuracy means and standard deviations over three datasets and ten seeds. EWF-Fuse exceeds probability averaging, voting, and raw Dempster–Shafer combination at all four displayed corruption levels; it is itself learned from labels. At 30%, its mean accuracy is 90.2 ± 8.5%, versus 77.6 ± 25.4%, 84.7 ± 15.1%, and 82.9 ± 12.0% for those comparators. Stacking reaches 93.3 ± 4.9%. The displayed clean-to-30% changes are approximately 4.1 percentage points for EWF-Fuse, 8.5 for raw Dempster–Shafer, 7.5 for averaging, and 2.3 for stacking. These differences use rounded table means and may differ from unrounded run-level contrasts. Pooled standard deviations reflect dataset heterogeneity as well as repeated-split variability; they are not confidence intervals.

**Table 1. Mean test accuracy (%) +/- standard deviation over 30 runs (3 datasets x 10 seeds) per corruption level c of source 2.**

| Method | c = 0% | c = 10% | c = 20% | c = 30% |
|---|---|---|---|---|
| Probability averaging | 92.2 +/- 6.7 | 89.7 +/- 9.9 | 87.1 +/- 12.5 | 84.7 +/- 15.1 |
| Majority vote | 88.9 +/- 8.4 | 87.5 +/- 9.0 | 85.4 +/- 10.3 | 82.9 +/- 12.0 |
| Stacking | 95.6 +/- 2.4 | 94.9 +/- 3.0 | 93.7 +/- 3.5 | 93.3 +/- 4.9 |
| Dempster-Shafer | 86.1 +/- 14.1 | 80.4 +/- 22.5 | 78.4 +/- 24.4 | 77.6 +/- 25.4 |
| EWF-Fuse (ours) | 94.3 +/- 3.5 | 93.8 +/- 4.1 | 91.4 +/- 6.1 | 90.2 +/- 8.5 |




a.




b.

**Fig. 1. Robustness evaluation of the five fusion methods under increasing source-2 corruption; a. Aggregate test accuracy versus source-2 corruption level for the five fusion methods (error bars: +/-1 standard deviation over 30 runs), and b. Robustness cost per dataset: Accuracy drop from c = 0% to c = 30% in percentage points.**

The source reports a 93.2–96.4% breast-cancer range across the grid and a 97.6% tie between EWF-Fuse and stacking on wine at 30% corruption. On digits, raw Dempster–Shafer accuracy changes from 66.7% to 42.6%, while EWF-Fuse changes from 90.4% to 78.8% and stacking from 93.7% to 87.4%. These observations are consistent with sensitivity of the tested combination and classifiers to corruption. They do not isolate conflict normalization as the sole cause, because evidence construction, source accuracy, and confidence change together.

## 5.3 | Ablation and Sensitivity

*Table 2* compares the full supervised realization with its ablations. Uniform weights reduce aggregate accuracy at 30% corruption from 90.2% to 78.0%, with digits accuracy changing from 78.8% to 43.2%. This supports the usefulness of validation-based weights within this implementation; it does not test the unsupervised estimator or prove the corrupt source is always identified. Weight exponents 1 and 3 yield aggregate accuracies 87.8% and 89.7%, respectively, against 90.2% for exponent 2.

Removing consensus adaptation changes displayed aggregate accuracy from 91.4% to 91.2% at 20% corruption and from 90.2% to 90.0% at 30%, a 0.2-point difference in both cases. Removing the conflict fallback leaves the displayed accuracy unchanged. These small ablation differences do not establish an unmeasured benefit on longer streams. The representative logged run shows the source-2 effective strength decreasing from about 0.32 to 0.19 over ten batches.

**Table 2. Ablation of EWF-Fuse: Aggregate accuracy (%) over 30 runs and digits-only accuracy (%) over 10 seeds at c = 20% and c = 30%.**

| Variant | Agg. c = 20% | Agg. c = 30% | Digits c = 20% | Digits c = 30% |
|---|---|---|---|---|
| EWF-Fuse (full, beta = 2) | 91.4 | 90.2 | 83.5 | 78.8 |
| no provenance discounting (no EWMA) | 91.2 | 90.0 | 82.8 | 78.8 |
| no conflict-adaptive fallback | 91.4 | 90.2 | 83.5 | 78.8 |
| uniform weights (1/3 each) | 79.9 | 78.0 | 48.6 | 43.2 |
| beta = 1 | 89.7 | 87.8 | 78.1 | 72.3 |
| beta = 3 | 91.1 | 89.7 | 83.3 | 78.6 |

## 5.4 | Robustness and Stress Behavior

*Fig. 2* shows a reported digits run at 30% corruption, seed 0. Per-batch mean raw conflict is approximately 0.30–0.37, while the fraction of instances using the fallback is approximately 0.07–0.20. A batch mean below $\tau = 0.5$ can coexist with individual instances above $\tau$; the branch must therefore be interpreted at the instance level. The source reports decreasing mean conflict as corruption increases because flatter outputs reduce entropy-based strength. The effective source coefficients in panel b) do not sum to one and should be read as discounted strengths, separately from normalized averaging weights $\beta$. Declining agreement with a consensus is not itself proof of corrupted-source identification.

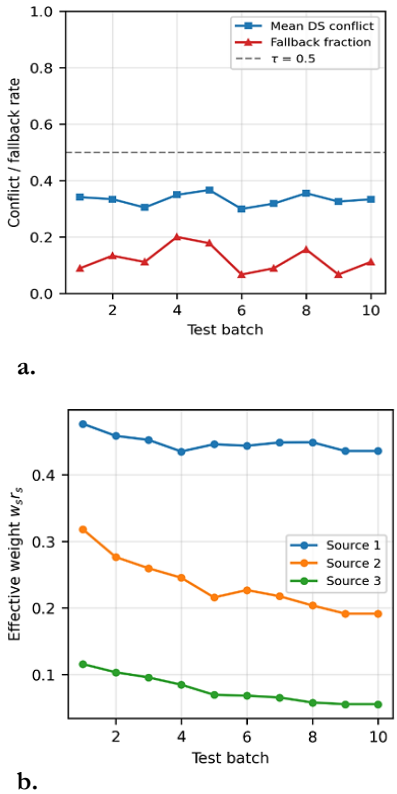



**Fig. 2. Reported digits run with c = 30% and seed 0. Panel; a. shows per-batch mean raw pairwise conflict and the fraction of instances exceeding the fallback threshold, and Panel b. shows effective source strengths, which are distinct from normalized β weights.**

At 20% corruption, voting has the lowest displayed ECE (0.090), and stacking has the lowest ECE among the continuous soft combiners (0.127) and the lowest NLL (0.29). EWF-Fuse has the highest ECE (0.319) and NLL 0.63. Raw Dempster–Shafer has the highest NLL (2.02), but its ECE (0.196) is below EWF-Fuse's. Discounting can flatten decision probabilities, yet unsigned ECE does not establish systematic

underconfidence; reliability diagrams or signed calibration summaries would be needed. Any recalibration must use a separate training or validation partition.

**Table 3. Aggregate accuracy (%), ECE, and NLL at c = 20% (mean +/- standard deviation over 30 runs).**

| Method | Accuracy (%) | ECE | NLL |
|---|---|---|---|
| Probability averaging | 87.1 +/- 12.5 | 0.179 +/- 0.056 | 0.47 +/- 0.29 |
| Majority vote | 85.4 +/- 10.3 | 0.090 +/- 0.046 | 1.18 +/- 1.09 |
| Stacking | 93.7 +/- 3.5 | 0.127 +/- 0.057 | 0.29 +/- 0.14 |
| Dempster-shafer | 78.4 +/- 24.4 | 0.196 +/- 0.211 | 2.02 +/- 2.50 |
| EWF-fuse (ours) | 91.4 +/- 6.1 | 0.319 +/- 0.106 | 0.63 +/- 0.36 |

## 5.5|Statistical Analysis

The source applies Friedman and paired Wilcoxon tests to 30 dataset-by-seed blocks per condition, with Holm adjustment across four EWF-Fuse comparisons [23], [24]. At 30% corruption it reports Friedman $\chi^2(4) = 54.4$, $p = 4.2\times10^{-11}$ and the contrasts in *Table 4*. These are reported exploratory statistics: repeated random splits overlap within only three datasets, so the 30 blocks are not 30 independent datasets. The p-values do not by themselves validate broad population significance. Dataset-level or hierarchical, dependence-aware analysis and raw paired outcomes are needed. The reported NLL comparison at 20% gives $W = 0$, $p = 1.9\times10^{-9}$, subject to the same dependence limitation.

**Table 4. Paired Wilcoxon signed-rank tests of EWF-Fuse against each baseline on test accuracy at c = 30% (n = 30 paired dataset-by-seed blocks), with Holm-adjusted p-values over the four comparisons, and the Friedman test across all five methods. Delta is the accuracy difference (EWF-Fuse minus baseline) in percentage points.**

| Comparison | W | p | p (Holm) | Delta (pp) |
|---|---|---|---|---|
| Friedman, all 5 methods | chi2(4) = 54.4 | 4.2e-11 | - | - |
| EWF-fuse vs probability averaging | 19 | 0.0005 | 0.0010 | +5.5 |
| EWF-fuse vs majority vote | 5 | 6.4e-06 | 2.6e-05 | +7.2 |
| EWF-fuse vs stacking | 13 | 0.0016 | 0.0016 | -3.1 |
| EWF-fuse vs dempster-shafer | 0 | 0.0002 | 0.0006 | +12.5 |

## 5.6|Discussion Findings, Losses, and Scope

The reported campaign shows smaller pooled accuracy loss under the tested corruption protocol and a substantial advantage of supervised weights over uniform weights. The evidence does not evaluate agreement-based unsupervised convergence or its error guarantees. The stated byte-identical reruns address deterministic output in one environment only, and do not establish complete replay or negligible provenance overhead.

Stacking has the highest pooled accuracy at every corruption level and leads EWF-Fuse by the reported paired contrast of 3.1 points at 30% corruption (Holm-adjusted $p = 0.0016$, with the dependence limitation above). Both methods have access to labelled validation information. A meta-learner can use class-specific patterns that scalar reliability weights cannot, but this mechanism is an interpretation rather than an isolated causal finding. The wine tie does not establish that small validation sets generally favor EWF-Fuse. Reliability-only deployment and label-free operation remain separate untested questions.

The corruption protocol modifies only source 2 through coupled dropout and training-label noise. Multi-source, adversarial, and drifting corruption are untested. Reliability is learned from validation labels rather than the conceptual agreement estimator. Test-batch adaptation also changes the inference protocol; confirmation that each batch is scored before updates and that no test labels enter any update is needed to assess causal online use. The calibration deficit limits probability-sensitive deployment without further validation.

The reported outcomes are proof-of-concept observations on three tabular benchmarks. Independent execution, dataset-level inference, and evaluation on genuinely heterogeneous acquisition sources are needed before stronger application claims can be made.

# 6 | Discussion Limitations and Counterarguments

## 6.1 | Limitations

The conceptual consonant mapping assumes possibility semantics and does not capture arbitrary non-consonant evidence. Agreement cannot identify coherent collusion without additional assumptions or anchors [15]. Repeated combination may enlarge the focal-set space and sharpen dependent evidence. The supervised campaign does not validate the unsupervised estimator, complete provenance architecture, or actual sensor/database/human fusion. Its repeated splits overlap and its pooled metrics mix three datasets. Code, raw outcomes, exact mass allocation, adaptation order, and independent replay remain material evidence requirements.

## 6.2 | Alternative Viewpoints and Responses

Alternative probability, belief-function, and possibility models should be compared under the same input semantics and validation access. The present campaign demonstrates the value of supervised reliability weighting within one implementation while stacking remains stronger in pooled accuracy. It does not establish an advantage when validation labels are unavailable. Provenance requirements can be applied to any of these calculi, and their implementation should be tested independently of the fusion-score comparison.

# 7 | Conclusion

The reported supervised EWF-Fuse realization retains 90.2 ± 8.5% pooled accuracy at 30% source-2 corruption, compared with 77.6%, 84.7%, and 82.9% for raw Dempster–Shafer, averaging, and voting. Uniform weighting lowers its accuracy to 78.0%, making reliability estimation the strongest observed ablation effect. These results concern constructed classifier sources and an overlapping repeated-split campaign; the retained statistical contrasts are exploratory.

Stacking remains ahead in pooled accuracy, and EWF-Fuse has the highest displayed ECE at 20% corruption. Future work should implement and validate the agreement-based estimator, publish the exact mass and update rules, separate calibration data from evaluation, quantify dependence-aware uncertainty, and audit complete provenance replay. The distinction between the conceptual consonant architecture and the tested entropy-strength realization must remain explicit when extending the method.

# Author Contributions

R. H.: Study design, methodology, and original drafting. S. Z.: Implementation, experimental execution, and data analysis. M. A.: Interpretation, supervision, and writing review and editing.

# Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

# Data Availability

The reported datasets are publicly accessible through scikit-learn [18], [19]. The supplied manuscript does not establish an archived study-specific dataset package. Input versions, splits, and corruption draws are required to reproduce the exact campaign.

# Code Availability

The source identifies scripts/exp/exp_16.py, scripts/results/exp_16_results.json, and two generated figures as supporting files. An accessible archive or supplied executable package has not been established by the manuscript. Exact code, environment, seeds, and raw outputs are required for independent reproduction.

# Generative AI Disclosure

The source reports AI-assisted drafting under the authors' direction and scientific responsibility.

# Conflicts of Interest

The authors declare no conflicts of interest.

# References

[1] Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, *85*(1), 6–23. https://doi.org/10.1109/5.554205

[2] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338–353. http://dx.doi.org/10.1016/S0019-9958(65)90241-X

[3] Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, *38*(2), 325–339. https://doi.org/10.1214/aoms/1177698950

[4] Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press. https://www.glennshafer.com/books/amte.html

[5] Yin, X., Han, J., & Yu, P. S. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE trans. on knowl. and data eng.*, *20*(6), 796–808. https://doi.org/10.1109/TKDE.2007.190745

[6] Zadeh, L. A. (1986). A simple view of the Dempster Shafer theory of evidence and its implication for the rule of combination. *AI magazine*, *7*(2), 85. https://doi.org/10.1609/aimag.v7i2.542

[7] Yager, R. R. (1987). On the Dempster Shafer framework and new combination rules. *Information sciences*, *41*(2), 93–137. https://doi.org/10.1016/0020-0255(87)90007-7

[8] Murphy, C. K. (2000). Combining belief functions when evidence conflicts. *Decision support systems*, *29*(1), 1–9. https://doi.org/10.1016/S0167-9236(99)00084-6

[9] Smets, P., & Kennes, R. (1994). The transferable belief model. *Artificial intelligence*, *66*(2), 191–234. https://doi.org/10.1016/0004-3702(94)90026-4

[10] Mercier, D., Quost, B., & Denœux, T. (2008). Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information fusion*, *9*(2), 246–258. https://doi.org/10.1016/j.inffus.2006.08.001

[11] Dubois, D., & Prade, H. (1988). Representation and combination of uncertainty with belief functions and possibility measures. *Computational intelligence*, *4*(3), 244–264. https://doi.org/10.1111/j.1467-8640.1988.tb00279.x

[12] Edalatpanah, S. A. (2026). The awareness logic of ambiguity (ALA): Triadic interpretive states and their structure-preserving operational core. *Journal of fuzzy extension and applications*, *7*(2), 661–702. https://doi.org/10.22105/jfea.2026.582827.2294

[13] Denœux, T. (2008). Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artificial intelligence*, *172*(2–3), 234–264. https://doi.org/10.1016/j.artint.2007.05.008

[14] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., … & Han, J. (2016). A survey on truth discovery. *ACM sigkdd explorations newsletter*, *17*(2), 1–16. https://doi.org/10.1145/2897350.2897352

[15] Dong, X. L., Srivastava, D., & Berti Équille, L. (2009). Integrating conflicting data. *Proceedings of the Vldb Endowment*, *2*(1), 550–561. https://www.vldb.org/pvldb/vol2/vldb09-pvldb47.pdf

[16] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., … & Van den Bussche, J. (2011). The open provenance model core specification (v1. 1). *Future generation computer systems*, *27*(6), 743–756. https://doi.org/10.1016/j.future.2010.07.005

[17] Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of machine learning research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[19] Developers, S. (n.d.). *Toy datasets*. https://scikitlearn.org/stable/datasets/toy_dataset.html

[20] Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, *20*(3), 226–239. https://doi.org/10.1109/34.667881

[21] Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[22] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International conference on machine learning* (pp. 1321–1330). Proceedings of Machine Learning Research (PMLR). https://proceedings.mlr.press/v70/guo17a.html

[23] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, *7*, 1–30. https://www.jmlr.org/papers/v7/demsar06a.html

[24] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733