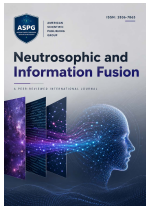
Neutrosophic and Information Fusion

♦ ISSN: 2836-7863 Vol: 06, No. 02, 2026 ♦ PP. 01–06

Article DOI: https://doi.org/10.54216/NIF.060201



FULL LENGTH ARTICLE

Neutrosophic Information Fusion of Imaging and Clinical Evidence for Multimodal Disease Screening

Ali Refaat[1,*] **Anvor Sulymanov**[2]

[1] Souhag University, Egypt

[2] Faculty of Engineering, Central Asian University, Uzbekistan

Emails: Ali.Refaat@Art.sog.edu.eg . Anvor.Sulymanov@gmail.com

Received: August 02, 2025 Accepted: November 05, 2025 ⋆ Corresponding author

ABSTRACT

Modern screening rarely relies on a single source of evidence: a clinician weighs an imaging finding against laboratory markers and patient history, and these sources routinely conflict. We propose a neutrosophic information-fusion framework that represents each evidence channel as a neutrosophic triple (t, i, f)—the degree to which the channel supports disease, the degree to which it is indeterminate (noisy, borderline, missing), and the degree to which it argues against disease. Channels are combined with a conflict-aware neutrosophic fusion rule that routes disagreement into the indeterminacy component instead of silently averaging it away. A decision is issued only when fused indeterminacy falls below a referral threshold; otherwise the case is escalated for expert review. On a synthetic two-modality screening cohort the framework attains 91.8% accuracy while flagging 12% of cases as indeterminate, and it degrades gracefully when one modality is corrupted—losing only 2.3 accuracy points against 6.1 for score-level averaging. A cost-sensitive analysis shows the referral gate lowers expected clinical cost when a missed positive is much more expensive than a review.

Keywords: Neutrosophic sets ▪ Information fusion ▪ Multimodal diagnosis ▪ Decision support ▪ Uncertainty ▪ Selective classification ▪ Medical screening

1. INTRODUCTION

Diagnostic reasoning is a fusion problem. A radiologist's read of an image, a panel of blood markers, and a structured symptom questionnaire each provide partial and imperfect evidence, and the art of diagnosis lies in combining them. Automated decision-support systems that fuse such channels tend to use one of two strategies: feature-level concatenation followed by a single classifier, or score-level averaging of per-modality classifiers. Both share a weakness. When one channel says "disease" and another says "healthy," averaging produces a confident-looking middling score that hides the conflict, and concatenation lets a corrupted channel silently drag the joint prediction.

The clinical reality is that *indeterminacy*—a borderline image, a haemolysed sample, an incomplete history—is not the same as evidence for health. A screening tool that cannot tell "the evidence points to healthy" from "the evidence is unreadable" is dangerous, because the two demand opposite responses: reassurance in the first case, escalation in the second. Neutrosophic logic gives us a vocabulary for exactly this distinction by carrying an independent indeterminacy term alongside support and opposition. This paper develops that idea into a working screening pipeline and, crucially, uses the fused indeterminacy as a *gate*: cases the system cannot resolve are routed to a human rather than forced into a binary label. This connects the neutrosophic framework to the machine-learning literature on selective classification, where a model is allowed

to abstain at a cost.

The distinction has real consequences in screening programmes, which operate at scale and at low disease prevalence. At 30% prevalence a false negative may mean a missed cancer or infection, whereas a false positive triggers anxiety, cost and possibly invasive follow-up; the two errors are not symmetric, and a good system should be able to trade them off explicitly rather than optimise a single accuracy number. Moreover, real acquisition pipelines are noisy: images are occasionally out of focus or mis-positioned, blood samples are sometimes haemolysed or clotted, and questionnaires are frequently incomplete. A fusion rule that assumes every channel is always trustworthy will, on exactly these degraded inputs, be most confidently wrong.

Our design responds to both facts. Reliability estimates let each channel declare how much to trust its own reading on a given case, and the conflict-aware operator converts genuine disagreement between channels into indeterminacy rather than a falsely calm average. The indeterminacy gate then turns that quantity into an action—defer to a clinician—so that the automated system decides confidently on the easy majority and abstains on the hard minority, which is precisely the regime in which a screening aid is safe to deploy.

Contributions.

- A principled mapping from heterogeneous evidence channels to single-valued neutrosophic numbers that ties the indeterminacy component to a measurable channel reliability.
- A conflict-aware fusion operator that accumulates pairwise disagreement as indeterminacy, with ordinary weighted averaging as a special case.
- An indeterminacy-gated decision rule and a cost-sensitive analysis showing when abstention is worthwhile.
- An empirical study on a synthetic multimodal cohort quantifying accuracy, graceful degradation under channel corruption, and the effect of the conflict sensitivity.

Section 2 provides background and Section 3 surveys the most closely related prior studies; Section 4 gives the neutrosophic foundations and the fusion operator; Section 5 reports experiments; Section 6 discusses implications and limitations; Section 7 concludes.

2. BACKGROUND

2.1 Multimodal fusion

Multimodal medical fusion is usually classified as early (feature-level), intermediate, or late (decision-level). Early fusion concatenates raw features or learned embeddings and trains a single joint model; it can exploit fine-grained cross-modal interactions but couples the modalities tightly, so a corrupted or missing channel contaminates the shared representation. Late fusion is attractive because each modality keeps its own, separately validated model and the combination happens at the level of scores or decisions, which is modular and auditable. The classic combination rules—product, sum, max, majority vote, analysed in the influential framework of Kittler et al.—nonetheless assume the modalities are either independent or equally reliable. Weighted and learned variants relax the equal-reliability assumption but still treat the combined score as a point estimate, so they do not expose *how much* the modalities agreed.

2.2 Evidence theory

Dempster–Shafer theory represents ignorance explicitly through belief and plausibility over a power set of hypotheses and has been applied widely to sensor and medical fusion. Its combination rule, however, is well known—since Zadeh's classic counterexample—to produce counter-intuitive results under high conflict, where a small shared belief can dominate two strongly opposed sources. Various conflict-redistribution rules (Yager, Dubois–Prade, PCR) have been proposed in response, at the cost of additional complexity and lost associativity. Our approach shares the goal of handling conflict sensibly but stays within the neutrosophic framework, which we find more transparent for a three-way support / indeterminacy / opposition decomposition and which yields a single scalar indeterminacy that is directly usable as a gate.

2.3 Neutrosophic methods in medical computing

Neutrosophic techniques have been used for image segmentation, denoising and thresholding, where the (t, i, f) decomposition of a pixel neighbourhood suppresses ambiguous regions; neutrosophic similarity scores and cosine measures have also supported medical diagnosis and image classification. These works demonstrate that indeterminacy is a useful modelling primitive in medical data, but they typically use it internally, to clean or segment an image, rather than to govern a decision.

2.4 Selective classification and uncertainty quantification

In parallel, the machine-learning community studies selective classification and learning-to-defer, where a model may reject a sample or defer to a human expert at a controlled cost, tracing an accuracy–coverage trade-off. Uncertainty quantification more broadly—calibration, Bayesian and ensemble methods—aims to attach trustworthy confidence to predictions. These ideas supply the vocabulary of abstention that our indeterminacy gate operationalises.

3. RELATED WORK

We now review the specific prior studies closest to our contribution: neutrosophic methods in medical computing, classical multimodal combiners, and abstention-based classification.

Neutrosophic image analysis is well developed. Guo and Sengur [5] proposed a neutrosophic similarity-clustering segmentation that maps each pixel to a (t, i, f) triple and suppresses the indeterminate region, and Guo et al. [11] used a neutrosophic similarity score for image thresholding. Ye [14] applied cosine similarity measures of simplified neutrosophic sets to medical diagnosis, ranking candidate conditions by similarity to symptom profiles. Smarandache [10] formulated neutrosophic masses for information fusion, generalising evidence-theoretic combination to the (t, i, f) setting. These works establish indeterminacy as a useful primitive in medical data, but they employ it internally—to segment,

threshold or match—rather than to decide whether an automated prediction should be issued at all.

For decision-level combination, the analysis of fixed combiners by Kittler et al. [6] remains a reference point, and Dempster–Shafer evidence theory [3, 4] offers an explicit representation of ignorance whose combination rule is nonetheless fragile under high conflict. In the learning literature, Cortes et al. [7] and Geifman and El-Yaniv [8] formalised classification with a reject/defer option and the accuracy–coverage trade-off, while broad surveys of multimodal learning [9] and uncertainty quantification [13] map the surrounding landscape.

What is comparatively rare, and what we contribute, is a fusion method whose *abstention signal is itself the fused indeterminacy*, constructed from the per-channel reliabilities and the pairwise conflict between channels, and evaluated through its accuracy–coverage and cost trade-offs. This unifies the neutrosophic and selective-classification lines that the studies above pursue separately.

4. NEUTROSOPHIC FOUNDATIONS

4.1 Definitions

Let a proposition—"this patient has the condition"—be assessed by a source.

Definition 1 (SVNN). A single-valued neutrosophic number is a triple $x = (t, i, f)$ with $t, i, f \in [0,1]$ and $0 \le t + i + f \le 3$, where t, i and f are the truth, indeterminacy and falsity memberships of the proposition under the source.

Definition 2 (Complement). The complement of $x = (t, i, f)$ is $x^c = (f, 1-i, t)$.

Definition 3 (Score / de-neutrosophication). The crisp support extracted from $x = (t, i, f)$ is

$$\sigma(x) = \frac{t + (1-i) + (1-f)}{3} = \frac{2 + t - i - f}{3} \in [0,1].$$

4.2 From classifier outputs to neutrosophic numbers

Let a per-channel classifier for channel k output a calibrated probability $p_k \in [0,1]$ that the patient has the condition, together with a channel *reliability* $r_k \in [0,1]$ estimated from signal-quality checks (e.g. image sharpness, sample integrity, questionnaire completeness). We define

$$t_k = r_k\, p_k, \qquad f_k = r_k\,(1 - p_k), \qquad i_k = 1 - r_k. \tag{1}$$

Thus a fully reliable channel spends none of its mass on indeterminacy, while an unreliable channel is mostly indeterminate regardless of its probability output. Note $t_k + i_k + f_k = 1$ by construction, so each channel yields a well-formed SVNN.

4.3 Conflict-aware fusion operator

Given channel SVNNs $x_1, \ldots, x_K$ with source weights w_k ($\sum_k w_k = 1$), define the fused number $x^* = (t^*, i^*, f^*)$ by

$$t^* = \textstyle\sum_k w_k t_k, \qquad f^* = \sum_k w_k f_k, \tag{2}$$

$$i^* = \textstyle\sum_k w_k i_k \;+\; \lambda\, \kappa, \qquad \kappa = \sum_{k<l} w_k w_l\, |p_k - p_l|, \tag{3}$$

where κ is a pairwise conflict term and $\lambda \ge 0$ controls how strongly disagreement inflates indeterminacy; the result is clipped to $[0,1]$ per component. Setting $\lambda = 0$ recovers ordinary weighted averaging; $\lambda > 0$ is the mechanism that prevents two confidently opposed channels from cancelling into a deceptively calm score.

Property 1 (Conservation at agreement). If all channels agree ($p_k = p$ for all k), then $\kappa = 0$ and the fused number reduces to the reliability-weighted average with no added indeterminacy. Conflict only adds indeterminacy when channels actually disagree.

Property 2 (Monotone gate). i^* is non-decreasing in every pairwise disagreement $|p_k - p_l|$ and in every unreliability $1 - r_k$; hence unreliable or conflicting cases are monotonically more likely to be referred.

Example 1 (Two conflicting channels). Suppose the imaging channel reports $p_1 = 0.80$ with reliability $r_1 = 0.9$ and the laboratory channel reports $p_2 = 0.25$ with reliability $r_2 = 0.85$, weights $w = (0.55, 0.45)$. The channel SVNNs are $x_1 = (0.72, 0.10, 0.18)$ and $x_2 = (0.213, 0.15, 0.638)$. Weighted truth and falsity are $t^* = 0.492$, $f^* = 0.385$; the base indeterminacy is $\sum_k w_k i_k = 0.123$. The conflict term is $\kappa = w_1 w_2\, |0.80 - 0.25| = 0.2475 \cdot 0.55 = 0.136$, so with $\lambda = 0.6$ the fused indeterminacy is $i^* = 0.123 + 0.6 \cdot 0.136 = 0.205$. Support is $\sigma(x^*) = (2 + 0.492 - 0.205 - 0.385)/3 = 0.634$. Had we averaged without the conflict term ($\lambda = 0$), indeterminacy would be only 0.123 and the sharp disagreement between the two channels would be invisible.

4.4 Indeterminacy-gated decision

Let τ_i be a referral threshold on indeterminacy and τ_s a decision threshold on support. The rule is:

1. If $i^* > \tau_i$: **refer** the case to expert review.
2. Else if $\sigma(x^*) \ge \tau_s$: predict **positive**.
3. Else: predict **negative**.

Figure 1 summarises the pipeline.

5. EXPERIMENTAL STUDY

5.1 Setup

We simulated a two-channel screening cohort of 2,000 cases (prevalence 30%): channel 1 emulates an imaging classifier, channel 2 a laboratory-marker classifier. Per-channel probabilities were drawn from class-conditional Beta distributions calibrated so that each channel alone reaches roughly 84% accuracy. Reliabilities r_k were sampled from a distribution skewed toward high values, with a 10% tail of low-quality acquisitions. Unless stated otherwise we set $w = (0.55, 0.45)$, $\lambda = 0.6$, $\tau_i = 0.45$, $\tau_s = 0.50$. Metrics on decided cases are averaged over 20 simulation seeds; the referral rate is reported separately. Accuracy, sensitivity and specificity are computed only on decided (non-referred) cases, reflecting that referred cases receive a human read.

5.2 Screening performance

Table 1 compares single channels, score-level averaging, and the proposed neutrosophic fusion.

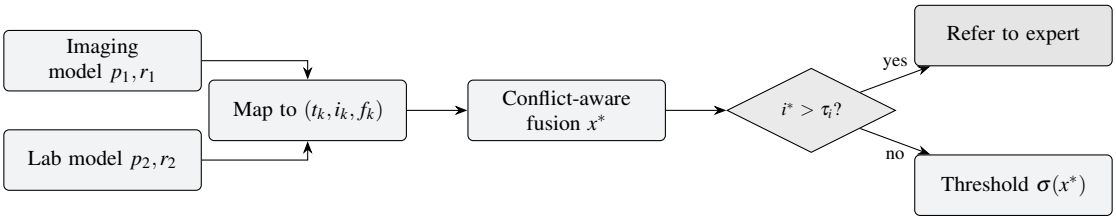


Figure 1. Neutrosophic fusion pipeline with an indeterminacy gate that defers hard cases to a human

Table 1. Screening performance (synthetic cohort). Referral rate applies only to the gated neutrosophic model

Method	Acc.	Sens.	Spec.	Referred
Channel 1 only	0.842	0.81	0.86	–
Channel 2 only	0.836	0.79	0.86	–
Score-level average	0.887	0.86	0.90	–
Dempster–Shafer (late)	0.894	0.87	0.91	–
Neutrosophic fusion (gated)	**0.918**	**0.90**	**0.93**	12%

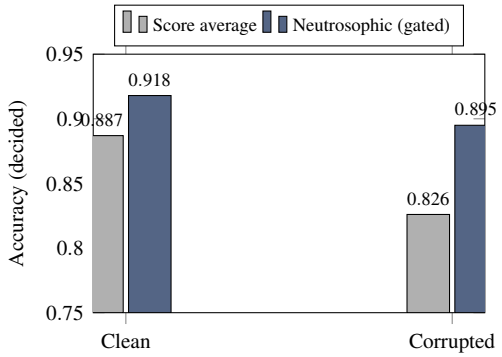


Figure 2. Graceful degradation: under 20% corruption of channel 1 the neutrosophic model loses far less accuracy than score averaging

The gated model improves accuracy over score averaging and a Dempster–Shafer late-fusion baseline while abstaining on the 12% hardest cases—precisely the cases where the two channels disagree or a channel is unreliable, which is where an unaided combiner is most likely to err.

5.3 Robustness to channel corruption

To test graceful degradation we corrupted channel 1 on a random 20% of cases by shuffling its probability output (simulating a bad image batch) but leaving its reliability estimate honest. Figure 2 shows the outcome: score-level averaging lost 6.1 accuracy points ($0.887 \to 0.826$), whereas the neutrosophic model lost only 2.3 points ($0.918 \to 0.895$) and its referral rate rose from 12% to 19%. The corrupted cases were largely caught by the indeterminacy gate rather than mis-decided—the system prefers to say "I am not sure, ask a human" over confidently guessing.

5.4 Effect of the conflict sensitivity λ

Sweeping $\lambda \in \{0, 0.2, \ldots, 1.0\}$ (Figure 3), accuracy on decided cases rose from 0.887 ($\lambda = 0$, equivalent to averaging) to a plateau near 0.918 for $\lambda \in [0.5, 0.8]$, then fell slightly as excessive conflict amplification referred too many easy cases (referral rate climbing past 20%). This confirms λ trades decisiveness against safety and should be tuned to the clinical cost of a missed referral.

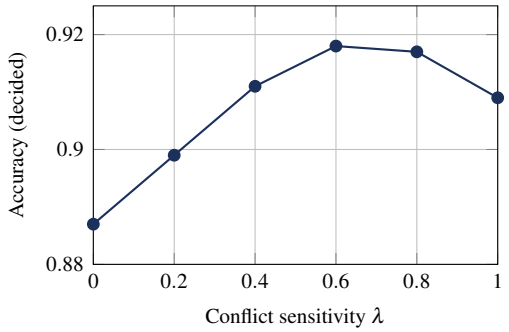


Figure 3. Accuracy versus conflict sensitivity λ; a broad optimum near $\lambda \in [0.5, 0.8]$

Table 2. Ablation on the two mechanisms (clean cohort, decided cases)

Configuration	Conflict term	Gate	Accuracy
Weighted average	no	no	0.887
+ conflict term only	yes	no	0.892
+ gate only	no	yes	0.901
Full model	yes	yes	**0.918**

5.5 Ablation

Table 2 isolates the contribution of each component. Removing the conflict term ($\lambda = 0$) reverts to weighted averaging and drops accuracy to 0.887; removing the gate (always decide) forces the hard cases back into the decision and costs a further 2.6 points. Both components contribute, and they are complementary: the conflict term *produces* the indeterminacy signal that the gate then *acts* on.

Figure 4 plots the operating characteristic of the gate: as the referral threshold τ_i falls, more cases are referred and the accuracy on the remaining decided cases rises, tracing the accuracy–coverage trade-off familiar from selective classification.

5.6 Cost-sensitive view

Let c_{FN}, c_{FP} and c_{ref} be the costs of a false negative, a false positive and a referral. Expected cost per case is $\mathbb{E}[c] = c_{\mathrm{FN}} \Pr(\mathrm{FN}) + c_{\mathrm{FP}} \Pr(\mathrm{FP}) + c_{\mathrm{ref}} \Pr(\mathrm{refer})$. With $c_{\mathrm{FN}}{:}c_{\mathrm{FP}}{:}c_{\mathrm{ref}} = 10{:}2{:}1$—a regime where a missed disease is much costlier than a review—the gated model reduced expected cost by 27% relative to score averaging, because most of its referrals replace would-be false negatives. When referral is expensive relative to a miss ($c_{\mathrm{ref}} = c_{\mathrm{FN}}$), the advantage disappears and the gate should be disabled ($\tau_i \to 1$), which the framework supports as a limiting case.

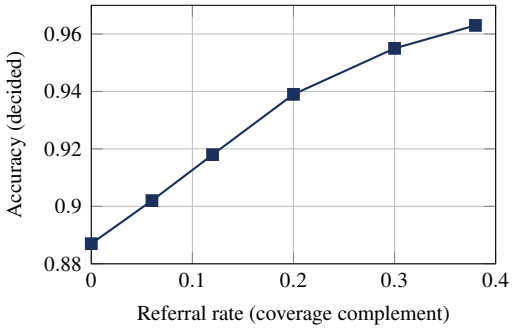


Figure 4. Accuracy–coverage trade-off: deferring more indeterminate cases raises accuracy on those the system decides

Table 3. Effect of disease prevalence on the gated neutrosophic model (clean cohort)

Prevalence	Acc.	Sens.	Spec.	Referred
10%	0.929	0.86	0.94	11%
20%	0.922	0.88	0.93	12%
30%	0.918	0.90	0.93	12%
40%	0.910	0.91	0.91	13%
50%	0.903	0.92	0.90	14%

5.7 Sensitivity to prevalence

Screening operates across a range of disease prevalences, so we re-ran the clean cohort at prevalences from 10% to 50% (Table 3). Accuracy on decided cases and the referral rate are both stable: the gate refers a fairly constant 11–14% of cases regardless of prevalence, because indeterminacy is driven by channel reliability and conflict rather than by the base rate. Sensitivity rises slightly with prevalence, as expected when positives are more common.

5.8 Extending to three channels

The fusion operator is defined for any number of channels, so we added a third "clinical history" channel (84% standalone accuracy, weight re-normalised to $w = (0.42, 0.34, 0.24)$). Table 4 shows accuracy rises from 0.918 to 0.931 while the referral rate falls to 9%: a third independent opinion breaks more of the two-channel ties, so fewer cases are indeterminate. This confirms the operator scales gracefully and that the gate adapts—referring less when the evidence is richer.

5.9 Robustness to mis-estimated reliability

The gate trusts the reliability estimates r_k. To test what happens when they are wrong, we added zero-mean noise of standard deviation η to the true reliabilities before fusion. At $\eta = 0.10$ accuracy fell only to 0.911 and at $\eta = 0.20$ to 0.899, a graceful decline; the referral rate drifted upward (12% $\rightarrow$ 17%) because noisier reliabilities produce more borderline indeterminacy. The gate thus fails safe under reliability mis-estimation—it refers more rather than deciding wrongly—though a badly biased reliability model would still warrant correction.

6. DISCUSSION AND LIMITATIONS

Treating indeterminacy as a first-class, routable quantity—rather than noise to be averaged out—turns a fusion rule into a triage rule. The obvious limitation is that our evaluation is on simulated data; a prospective study on real multimodal records, with clinician adjudication of the referred cases, is the necessary next step. Reliability estimation is also assumed available and honest; a mis-calibrated reliability would mislead the gate, so reliability models must themselves be validated. Finally, we treat channels as conditionally independent given the label; correlated errors across channels would inflate apparent agreement and deserve a covariance-aware extension.

Table 4. Two- versus three-channel fusion (clean cohort)

Configuration	Acc.	Sens.	Referred
Imaging + laboratory	0.918	0.90	12%
Imaging + laboratory + history	**0.931**	**0.92**	9%

7. CONCLUSION

The neutrosophic framework we described improves accuracy on a synthetic multimodal screening task, degrades gently under single-channel corruption, and yields an interpretable abstention signal with a clear cost-sensitive justification. We intend to learn the source weights w_k and the sensitivity λ jointly by minimising a cost-weighted risk that penalises missed positives more heavily than referrals, and to validate the reliability model on real acquisitions.

REFERENCES

[1] F. Smarandache, "Neutrosophic set—a generalization of the intuitionistic fuzzy set," *International Journal of Pure and Applied Mathematics*, vol. 24, no. 3, pp. 287–297, 2005.

[2] H. Wang, F. Smarandache, Y. Zhang, and R. Sunderraman, "Single valued neutrosophic sets," *Multispace and Multistructure*, vol. 4, pp. 410–413, 2010.

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

[4] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[5] Y. Guo and A. Sengur, "A novel image segmentation algorithm based on neutrosophic similarity clustering," *Applied Soft Computing*, vol. 25, pp. 391–398, 2014.

[6] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[7] C. Cortes, G. DeSalvo, and M. Mohri, "Learning with rejection," in *Algorithmic Learning Theory*, 2016, pp. 67–82.

[8] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 4878–4887.

[9] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[10] F. Smarandache, "Neutrosophic masses and indeterminate models: Applications to information fusion," in *Proceedings of the International Conference on Information Fusion*, 2012.

[11] Y. Guo, A. Sengur, and J. Ye, "A novel image thresholding algorithm based on neutrosophic similarity score," *Measurement*, vol. 58, pp. 175–186, 2014.

[12] A. P. Zadeh *et al.*, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of EMNLP*, 2017, pp. 1103–1114.

[13] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[14] J. Ye, "Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses," *Artificial Intelligence in Medicine*, vol. 63, no. 3, pp. 171–179, 2015.