

International Journal of Knowledge Processing Studies (KPS)



Homepage: <http://kps.artahub.ir/>



ORIGINAL RESEARCH ARTICLE

Explainable Multimodal Knowledge Processing for Digital Media Integrity Risk Assessment under Uncertainty

Rose Mary Mathew^{1,*}, Aqueena Joy

¹ Associate Professor, Department of Computer Applications, Federal Institute of Science and Technology, Kerala, India. rosem.mathew@gmail.com, 0000-0003-0555-4873.

² PG Student, Department of Computer Applications, Federal Institute of Science and Technology, Kerala, India.

ARTICLE INFO

Article History:

Received: 2026-06-24

Revised: 2026-07-22

Accepted: 2026-08-29

Published Online: 2026-09-01

Keywords:

Digital media integrity; Fuzzy inference; Misinformation detection; Multimodal learning; Risk assessment; Uncertainty.

Number of Reference: 25

Number of Figures: 3

Number of Tables: 14

DOI: 10.22034/kps.2026.599041.1304



ABSTRACT

The rapid growth of social media and generative artificial intelligence has increased the prevalence of misinformation and manipulated digital media, creating challenges for reliable content verification. This study presents a soft computing-based explainable multimodal framework for digital media integrity risk assessment under uncertainty. The framework integrates transformer-based textual credibility and emotion analysis, OCR-assisted text extraction, CLIP-based image-text semantic consistency, and ResNet50-based visual integrity assessment. The resulting evidence is integrated through a weight-adjusted Mamdani fuzzy inference mechanism to produce continuous integrity-risk scores and graded risk levels. SHAP and Grad-CAM provide supporting textual and visual explanations. On the benchmark test set, the framework achieved 95.6% accuracy and 95.1% F1-score, with a ROC-AUC of 0.97, indicating strong threshold-independent discrimination under the evaluated test conditions. Ablation and sensitivity analyses demonstrated performance changes associated with the integrated components and fusion configuration. Five repeated runs showed limited performance variation under the evaluated protocol. External validation on approximately 1,500 previously unseen multimodal samples achieved a 92.9% F1-score, indicating performance variation under a different data distribution. The framework provides an interpretable, uncertainty-aware approach to multimodal digital media integrity assessment and supports graded risk interpretation rather than relying solely on binary classification. ©authors.

► **Citation:** Rose Mary, M., & Joy, A. (2026). Explainable Multimodal Knowledge Processing for Digital Media Integrity Risk Assessment under Uncertainty. *International Journal of Knowledge Processing Studies (KPS)*, 6(3): 26-50. Doi: 10.22034/kps.2026.599041.1304

1. Introduction

The rapid growth of social media, online news platforms, and instant messaging has transformed the generation, distribution, and consumption of information. Although these technologies facilitate rapid communication and knowledge sharing, they have also accelerated the dissemination of misinformation, disinformation, manipulated media, and AI-generated content. The problem is particularly challenging when misleading information is distributed across textual, visual, and cross-modal forms (Vosoughi et al., 2018). Consequently, automated digital media verification has emerged as an important research problem spanning artificial intelligence, data mining, knowledge processing, and digital media forensics (Comito et al., 2023; Nakamura et al., 2020).

Early misinformation detection methods primarily relied on conventional machine-learning algorithms, including Support Vector Machines, Naïve Bayes, and decision trees, using handcrafted linguistic, stylistic, and statistical features. Although effective for constrained classification tasks, such approaches have limited ability to capture contextual semantics and long-range dependencies. The development of transformer-based language models substantially improved contextual representation learning. BERT, for example, demonstrated the effectiveness of bidirectional contextual representations for natural-language understanding and classification tasks (Devlin et al., 2019).

Misinformation has subsequently become increasingly multimodal. A misleading narrative may combine a false textual claim with an authentic but contextually unrelated image, or use an AI-generated or manipulated image as apparently supporting evidence. Multimodal datasets such as Fakeddit have therefore encouraged research beyond text-only misinformation detection (Nakamura et al., 2020). Recent studies have explored multimodal attention, transformer architectures, contrastive representation learning, and image-text consistency analysis

to exploit complementary evidence from different modalities (Comito et al., 2023).

Vision-language representation learning has further expanded the scope of multimodal knowledge processing. CLIP demonstrated that images and natural-language descriptions can be represented within a shared semantic space, enabling image-text similarity estimation (Radford et al., 2021). More recent studies have investigated vision-language models for multimodal misinformation detection and fact verification by combining textual and visual evidence, with some approaches additionally incorporating reasoning or external evidence retrieval (Tahmasebi et al., 2024; Jin et al., 2024). Recent benchmark research has also highlighted the diversity of multimodal misinformation, in which textual, visual, and cross-modal distortions may occur independently or jointly, posing challenges for general-purpose vision-language models (Liu et al., 2025).

Despite these developments, reliable digital media verification remains difficult because the available evidence may be incomplete, ambiguous, or contradictory. An authentic image may accompany a false caption, a valid claim may be expressed using emotionally manipulative language, or manipulated visual content may be associated with partially valid textual information. Under such conditions, a binary authentic/fake decision may not adequately represent the degree or nature of the available evidence. This motivates decision-support approaches capable of integrating heterogeneous evidence and representing intermediate or conflicting states rather than relying exclusively on rigid class labels (Jin et al., 2024; Yan et al., 2025).

To address this challenge, this study formulates digital media verification as an integrity-risk assessment problem under uncertainty. The proposed framework processes complementary evidence derived from textual credibility, emotional signals, text embedded in images, image-text semantic consistency, and visual integrity. These heterogeneous outputs are transformed into structured evidence variables and integrated through a Mamdani fuzzy

inference mechanism to generate a continuous integrity-risk assessment. The resulting risk level provides a graded representation of media integrity rather than restricting the final decision to a binary class. Explainability mechanisms are incorporated to provide supporting textual and visual evidence for interpretation and human verification.

The contribution of the study does not lie in introducing new BERT, CLIP, ResNet50, SHAP, or Grad-CAM architectures. Instead, the work addresses the knowledge-processing and decision-integration problem of transforming heterogeneous multimodal model outputs into an interpretable assessment under uncertainty. In particular, the fuzzy reasoning layer provides an explicit mechanism for combining evidence that may be uncertain or mutually inconsistent and for representing gradual risk states that may be obscured by conventional binary classification.

The main contributions of this study are:

1. Uncertainty-aware multimodal assessment: A unified framework is developed for digital media integrity assessment by jointly considering textual, emotional, OCR-derived, visual, and cross-modal evidence.
2. Fuzzy evidence integration: A Mamdani fuzzy inference mechanism is employed to transform heterogeneous evidence variables into continuous and interpretable integrity-risk levels, allowing intermediate and conflicting evidence states to be represented.
3. Multimodal knowledge integration: Textual credibility, emotional signals, embedded textual information, image-text semantic consistency, and visual integrity are processed as complementary evidence within a unified decision-support framework.
4. Explainable decision support: SHAP-based textual attribution and Grad-CAM-based visual evidence are incorporated to assist interpretation of the model outputs and support human-in-the-loop verification.
5. Empirical evaluation: The framework is evaluated using comparative experiments, repeated runs, statistical analysis, sensitivity and ablation studies, explainability analysis, and external validation to examine performance, robustness, and generalization.

The remainder of the paper is organized as follows. Section 2 presents the related literature and research background. Section 3 describes the proposed methodology and experimental setup. Section 4 presents and discusses the experimental results, including comparative evaluation, statistical analysis, sensitivity and ablation studies, explainability analysis, and external validation. Section 5 concludes the study and discusses future research directions.

2. Literature Review

Misinformation detection has evolved from handcrafted feature-based classification toward contextual representation learning and multimodal evidence analysis. Early approaches relied on statistical, linguistic, and stylistic features combined with conventional machine-learning classifiers. The emergence of transformer-based language models improved contextual representation learning by capturing bidirectional dependencies within text. BERT, for example, demonstrated the effectiveness of contextualized representations for natural-language understanding and classification tasks (Devlin et al., 2019). Nevertheless, text-only approaches are limited when the interpretation or credibility of a message depends on associated visual content.

The availability of multimodal datasets has consequently encouraged research that jointly considers textual and visual information. Fakeddit introduced a large-scale multimodal benchmark containing textual, visual, and contextual information for fake-news detection (Nakamura et al., 2020). A systematic survey by Comito et al. (2023) further documented the increasing use of visual representations, attention mechanisms, cross-modal interactions, and deep-learning architectures in multimodal misinformation research. These developments demonstrate a

shift from isolated modality analysis toward the joint processing of heterogeneous evidence.

Cross-modal semantic analysis is particularly relevant to out-of-context misinformation, in which an authentic image may be paired with an incorrect, outdated, or misleading textual claim. CLIP established a shared representation space for images and natural-language descriptions, providing a basis for estimating image–text semantic correspondence (Radford et al., 2021). Subsequent multimodal transformer approaches extended this direction by modelling interactions between textual and visual representations. However, semantic correspondence alone does not establish factual correctness. An image may be semantically related to a claim while still being presented in a misleading context; consequently, cross-modal similarity is more appropriately considered as one source of evidence within a broader verification process.

Visual forensic analysis provides another source of evidence by examining whether an image has been manipulated or synthetically generated. ResNet architectures established effective hierarchical visual representation learning and have subsequently influenced a range of image-analysis and manipulation-detection systems (He et al., 2016). Visual authenticity, however, is distinct from claim authenticity. An unaltered image may be used to support a false or misleading claim, while manipulated visual content may accompany text containing partially valid information. This distinction motivates the consideration of visual forensic evidence alongside semantic and contextual information rather than treating manipulation detection as a complete solution to misinformation assessment (Sagar et al., 2026).

The emergence of large vision-language models (LVLMs) has further expanded multimodal misinformation research. Tahmasebi et al. (2024) investigated multimodal misinformation detection using large vision-language models, while Jin et al. (2024) explored multimodal manipulation reasoning. More recent work, including TRUST-VL, has investigated explainable

multimodal misinformation detection across different types of content distortion (Yan et al., 2025). MMFakeBench further highlighted the diversity of multimodal misinformation by considering mixed-source textual, visual, and cross-modal distortions in the evaluation of LVLMs (Liu et al., 2025). Collectively, these studies demonstrate the growing role of multimodal reasoning and vision-language models in digital-media verification.

Despite these advances, an important challenge concerns the treatment of uncertain and conflicting multimodal evidence. Different evidence sources may provide inconsistent signals: textual content may appear credible while the associated image is manipulated, or an authentic image may be semantically inconsistent with its accompanying caption. Classification-oriented systems generally transform such evidence into discrete decisions, which can obscure intermediate credibility states. Similarly, fusion strategies based on concatenation, attention, voting, or fixed score combinations may not explicitly represent gradual changes in risk or provide an interpretable mechanism for reasoning when evidence sources disagree.

Explainability is another important consideration in multimodal decision support. SHAP provides feature-attribution information for interpreting model predictions, while Grad-CAM provides spatial evidence associated with visual predictions (Lundberg & Lee, 2017; Selvaraju et al., 2017). These methods can facilitate human inspection of textual and visual model outputs. However, explainability is frequently applied as a post-hoc interpretation layer, while comparatively less attention has been given to connecting explanatory evidence with multimodal evidence integration and uncertainty-aware decision processes. This creates an opportunity for decision-support frameworks in which model outputs and their associated explanations can be considered together during media-integrity assessment.

Soft-computing approaches provide a foundation for representing gradual and uncertain states. Fuzzy set theory allows observations to possess degrees of membership rather than requiring exclusive

assignment to discrete categories (Zadeh et al., 1965; Mamdani & Assilian et al., 1975). Fuzzy inference can consequently represent intermediate states such as moderate textual suspicion, partial image–text inconsistency, or uncertain visual manipulation. Such representations are relevant to digital-media verification because the credibility of content may depend on context and may not always correspond to a strictly binary authentic/fake state.

Based on the reviewed literature, four research gaps are identified. First, many existing misinformation-detection systems remain primarily classification-oriented and provide limited representation of intermediate integrity-risk states. Second, although multimodal methods integrate multiple sources of evidence, explicit reasoning over conflicting or uncertain evidence remains comparatively limited. Third, explainability is often treated primarily as a post-hoc interpretation mechanism rather than being considered alongside multimodal evidence integration. Fourth, recent LVLM-based approaches provide increasingly capable multimodal reasoning, but their computational requirements and dependence on large foundation models can present challenges for modular and interpretable decision-support applications (Tahmasebi et al., 2024; Liu et al., 2025; Yan et al., 2025).

The present study is positioned within these gaps from a knowledge-processing perspective. The focus is not on introducing a new BERT, CLIP, or ResNet architecture, but on the structured integration of heterogeneous evidence produced by established analytical models. Textual credibility, emotional signals, OCR-derived information, image–text consistency, and visual integrity are considered complementary evidence sources, while fuzzy reasoning is used to support continuous integrity-risk assessment. Explainability mechanisms provide additional textual and visual evidence for interpreting the resulting assessment. The resulting framework therefore treats digital-media verification as an evidence-integration and decision-support problem in which uncertainty and multimodal disagreement are explicitly considered.

The framework is evaluated using comparative experiments, repeated runs, confidence-interval analysis, fusion-weight sensitivity analysis, ablation experiments, and external validation. The evaluation examines predictive performance, robustness, generalization, and interpretability under the proposed multimodal evidence-processing approach. This positioning distinguishes the study from approaches that focus exclusively on binary misinformation classification and instead emphasizes the transformation of heterogeneous multimodal evidence into interpretable integrity-risk information.

3. Method

This study develops an explainable multimodal framework for digital media integrity-risk assessment under uncertainty. Rather than formulating media verification exclusively as a binary classification problem, the framework treats integrity assessment as a continuous risk-estimation task. This formulation is intended to accommodate heterogeneous, incomplete, and potentially conflicting evidence that may arise across textual and visual modalities.

For a media instance containing textual and/or visual content, the framework derives complementary evidence from five analytical sources: textual credibility, emotional signals, OCR-derived information, image–text semantic consistency, and visual integrity. Textual analysis estimates the credibility risk associated with the linguistic and claim-level characteristics of the input, while emotion analysis captures affective signals that may contribute to potentially manipulative communication. OCR extracts text embedded within visual content, such as memes and screenshots. Image–text analysis evaluates semantic correspondence between textual and visual information, and visual analysis estimates the likelihood of manipulated or synthetically generated content.

The outputs of these analytical components are treated as evidence variables rather than independent final decisions. They are subsequently transformed into a common risk representation and integrated through a Mamdani fuzzy inference mechanism. The

fuzzy reasoning layer represents gradual risk states and provides a mechanism for handling uncertainty and disagreement among heterogeneous evidence sources. The resulting continuous risk score is mapped to interpretable integrity-risk categories.

Explainability is incorporated to support the interpretation of the evidence contributing to the textual and visual assessments. The detailed mathematical formulations of the analytical components, evidence transformation, fusion mechanism, fuzzy inference process, and explainability procedures are presented in the subsequent subsections. The effect of the fusion parameters is examined through sensitivity analysis. The framework is designed as a research-oriented decision-support approach rather than as a claim of universal deployment readiness. Its effectiveness is evaluated using benchmark and external datasets, with emphasis on multimodal evidence integration, uncertainty-aware reasoning, interpretability, and generalization. The overall architecture is presented in Figure. 1.

3.1 System Architecture

The proposed framework adopts a modular pipeline for processing heterogeneous digital-media evidence. Let a media instance be represented as

$$M = \{T, I\}$$

where T denotes the textual content and I denotes the associated image content. The architecture supports text-only, image-only, and multimodal inputs according to the availability of the corresponding evidence sources.

The framework comprises the following six major layers:

1. Textual Credibility Analysis
2. Emotion Signal Analysis
3. OCR-Based Embedded Text Extraction
4. Image-Text Semantic Consistency Verification
5. Visual Integrity and Deepfake Detection
6. Soft Computing-Based Risk Fusion and Interpretation

The first five layers generate complementary evidence variables rather than

independent classification decisions. These variables are subsequently transformed into fuzzy linguistic representations and processed by the risk-inference layer to obtain a continuous integrity-risk score. The score is then mapped to predefined integrity-risk categories ranging from highly authentic to high fake risk.

The modular architecture separates evidence extraction from evidence integration, allowing information from different analytical sources to be retained before the final decision is produced. This organization is particularly relevant when evidence sources provide intermediate or conflicting signals. The explainability mechanisms are associated with the corresponding textual and visual analysis stages and provide supporting information for interpreting the resulting assessment.

The following subsections describe the individual analytical modules and their mathematical formulations, followed by the evidence-fusion mechanism, fuzzy risk inference, experimental configuration, and explainability procedure.

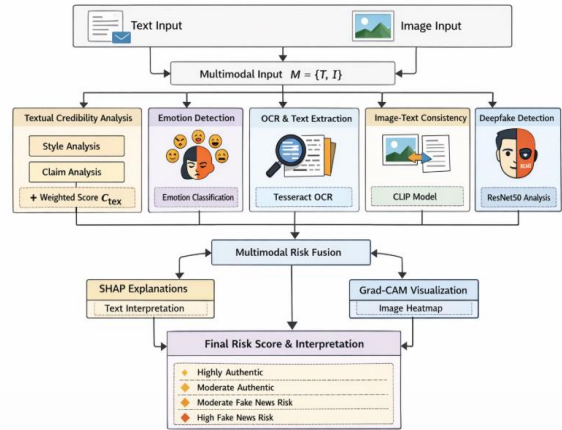


Figure 1. System architecture

3.2 Textual Credibility and Emotion Analysis

The textual analysis module derives two complementary evidence variables: textual credibility risk and emotion intensity. These variables are retained as intermediate evidence and are subsequently used by the multimodal risk-inference mechanism described in Section 3.6.

3.2.1 Linguistic Style and Factual Claim Pathway

The textual analysis employs two complementary BERT-based pathways. The

first estimates the probability of stylistic manipulation, denoted by P_{style} , while the second estimates the probability of factual inconsistency, denoted by P_{claim} . Combining these two signals allows linguistic and claim-level evidence to contribute jointly to the textual risk estimate.

The textual risk score is defined as

$$C_{text} = w_s P_{style} + w_c P_{claim}$$

subject to: $w_s + w_c = 1$

where w_s and w_c represent the style and claim weights, respectively. Based on validation experiments and sensitivity analysis, the optimal weight configuration was determined as:

$$w_s = 0.4, w_c = 0.6$$

Thus, the final textual credibility score becomes:

$$C_{text} = 0.4P_{style} + 0.6P_{claim}$$

The resulting score is normalized to [0,1] and represented using Low, Medium, and High fuzzy membership functions for subsequent risk inference.

3.2.2 Emotion Signal Detection

A BERT-based multi-label classifier estimates probabilities for emotions including anger, fear, joy, sadness, surprise, and neutral. (Demszyk et al., 2020). The emotion vector is represented as:

$$E = \{P_{anger} \cdot P_{fear} \cdot P_{joy} \cdot P_{sadness} \cdot P_{surprise} \cdot P_{neutral}\}$$

where P_i denotes the probability associated with the i^{th} emotion category. Since misinformation often contains subtle emotional signals hidden beneath apparently neutral language, a Neutral Suppression Filter (NSF) is introduced to reduce the dominance of the neutral class. The adjusted neutral probability is computed as:

$$P'_{neutral} = \lambda P_{neutral}$$

where λ is the suppression coefficient. Based on empirical validation, $\lambda = 0.3$ is selected. The emotion intensity score is then calculated as:

$$E_{score} = 1 - P'_{neutral}$$

The resulting emotion distribution provides an additional indicator of affective influence

and is subsequently incorporated into multimodal risk reasoning.

3.3 OCR-Based Text Extraction

Text embedded within memes, screenshots, and other visual media may provide additional evidence that is unavailable in the original textual input. To incorporate this information, Tesseract OCR is applied to the input image I to extract any machine-readable textual content. The extracted OCR text is represented as:

$$T_{ocr} = \text{OCR}(I)$$

where T_{ocr} denotes the text obtained from the image.

Since OCR output may contain recognition errors, fragmented characters, and formatting artifacts, a lightweight preprocessing procedure is applied before the extracted text is incorporated into the multimodal analysis. The preprocessing includes lowercase conversion, removal of punctuation and special symbols, elimination of non-alphanumeric noise, and removal of duplicate or spurious text fragments. The resulting cleaned representation is denoted by:

$$T_{ocr}^* = \text{Preprocess}(T_{ocr})$$

To minimize the influence of unreliable OCR fragments, text segments containing fewer than five characters are discarded. Thus, the validated OCR representation is defined as:

$$T_{ocr}^v = \{s \in T_{ocr}^* \mid |s| \geq 5\}$$

where s represents an extracted text segment and $|s|$ denotes its character length.

The validated OCR text is subsequently combined with the original textual input T associated with the media instance:

$$T^+ = T \oplus T_{ocr}^v$$

where $\oplus$ denotes text concatenation. This augmented textual representation T^+ enables information embedded within the visual content to participate in the subsequent textual credibility and emotion analysis described in Section 3.2. In this way, OCR serves as an information-recovery stage that connects visual content with the framework's textual evidence pathways without

introducing a separate credibility model for the extracted text.

3.4 Image–Text Consistency Verification

To assess potential contextual inconsistency between textual claims and their associated visual content, the framework employs CLIP-based image–text semantic verification. Given a text–image pair (T, I) , CLIP generates corresponding text and image embeddings:

$$z_T = f_T(T), z_I = f_I(I)$$

where $f_T(\cdot)$ and $f_I(\cdot)$ denote the CLIP text and image encoders, respectively.

The semantic relationship between the two modalities is measured using cosine similarity:

$$S_{\text{sim}} = \frac{z_T \cdot z_I}{\|z_T\| \|z_I\|}$$

Since cosine similarity lies in the interval $[-1, 1]$, it is first normalized to $[0, 1]$:

$$S_{\text{norm}} = \frac{S_{\text{sim}} + 1}{2}$$

The semantic mismatch evidence is then defined as

$$C_{\text{mismatch}} = 1 - S_{\text{norm}} = \frac{1 - S_{\text{sim}}}{2}$$

Thus, larger values of C_{mismatch} indicate weaker semantic correspondence between the textual and visual representations and therefore greater cross-modal inconsistency.

This normalized mismatch variable is passed to the subsequent risk-inference stage as one evidence source, rather than being interpreted independently as a misinformation decision.

3.5 Visual Integrity and Deepfake Detection

Visual integrity is assessed using a fine-tuned ResNet50 classifier trained to distinguish authentic images from synthetically manipulated images (Tolosana et al., 2020). For an input image I , the network produces two class logits, z_{real} and z_{fake} , corresponding to the authentic and manipulated classes. The manipulation probability is computed using the softmax function:

$$P_{\text{fake}} = \frac{e^{z_{\text{fake}}}}{e^{z_{\text{real}}} + e^{z_{\text{fake}}}}$$

Similarly, the probability of an authentic image is given by

$$P_{\text{real}} = \frac{e^{z_{\text{real}}}}{e^{z_{\text{real}}} + e^{z_{\text{fake}}}}$$

where $P_{\text{fake}} + P_{\text{real}} = 1$. The value P_{fake} is retained as the visual manipulation-risk evidence for the subsequent risk-inference stage.

To provide an interpretable indication of the image regions contributing to the prediction, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to the ResNet50 feature representations. For a target class c , the importance of feature map A^k is estimated from the gradients of the class score y^c with respect to the spatial activations:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where A_{ij}^k denotes the activation of feature map k at spatial location (i, j) , and Z represents the number of spatial locations in the feature map.

The Grad-CAM localization map is then obtained as

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right)$$

The resulting heatmap identifies image regions that contribute positively to the selected class prediction. For manipulated-image assessment, these regions provide visual evidence associated with the model's prediction and can be used to support interpretation of the automated assessment.

Unlike P_{fake} , which serves as a quantitative evidence variable for risk inference, the Grad-CAM output is used solely for visual explanation. This separation prevents the explanatory heatmap from being treated as an additional numerical prediction signal and avoids duplicating its role in the subsequent fusion process.

3.6 Multimodal Risk Fusion and Fuzzy Risk Interpretation

The four evidence variables, namely textual risk C_{text} , emotion risk E_{score} , image–text semantic mismatch C_{mismatch} , and visual manipulation probability P_{fake} , are integrated

using a weight-adjusted Mamdani fuzzy inference mechanism. The corresponding fusion weights are $w_1 = 0.40$, $w_2 = 0.10$, $w_3 = 0.25$, and $w_4 = 0.25$, satisfying $\sum_i w_i = 1$. These weights represent the relative contribution assigned to the four evidence sources during fuzzy rule activation.

Each evidence variable is fuzzified using three linguistic terms: Low, Medium, and High. For a rule k , the weighted firing strength is calculated as

$$\alpha_k = \min(w_1 \mu_{C_{text}}^{(k)}, w_2 \mu_E^{(k)}, w_3 \mu_{C_{mismatch}}^{(k)}, w_4 \mu_{P_{fake}}^{(k)})$$

where $\mu^{(k)}$ denotes the membership degree of the corresponding input in the linguistic state specified by rule k . Rules containing fewer than four antecedents use the corresponding subset of weighted membership terms. The resulting firing strength is used to truncate the consequent membership function:

$$\mu_k(r) = \min(\alpha_k, \mu_{B_k}(r))$$

where B_k denotes the output linguistic category associated with rule k . The activated consequents are aggregated using the maximum operator:

$$\mu_R(r) = \max_k \mu_k(r).$$

Finally, the continuous risk score is obtained through centroid defuzzification:

$$R_{final} = \frac{\int_0^1 r \mu_R(r) dr}{\int_0^1 \mu_R(r) dr}.$$

The resulting score is mapped to four graded risk states: Highly Authentic, Moderately Authentic, Moderate Fake Risk, and High Fake Risk according to the thresholds defined in this section. This formulation allows heterogeneous evidence sources to contribute with different relative weights while retaining partial membership and gradual transitions between neighbouring risk states.

3.6.1 Mamdani Fuzzy Inference System

The weight-adjusted Mamdani Fuzzy Inference System (FIS) serves as the risk-integration and interpretation layer. Unlike a direct numerical weighted sum, the fuzzy formulation represents gradual transitions among linguistic risk states and provides a rule-based mechanism for interpreting consistent, intermediate, and conflicting evidence across modalities.

The FIS receives the following four normalized inputs:

1. C_{text} — Textual Risk
2. E_{score} — Emotion
3. $C_{mismatch}$ — Semantic Mismatch
4. P_{fake} — Deepfake Probability

Each input is represented using three linguistic terms: Low (L), Medium (M), and High (H). Consequently, the theoretical input space contains $3^4 = 81$ possible linguistic combinations. The framework does not require a separate rule for every theoretical combination. Instead, a reduced expert-defined rule base is used to represent the principal and practically meaningful relationships among the four evidence variables. This design avoids redundant or marginal rules while retaining patterns associated with consistent, conflicting, and intermediate multimodal evidence.

3.6.2 Input Membership Functions

For an input variable x , a triangular membership function is defined as

$$\mu(x; a; b; c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right)$$

where a , b , and c denote the left endpoint, peak, and right endpoint, respectively.

All input membership functions operate over the normalized domain $[0, 1]$. The membership parameters are given in Table 1.

Table 1. Membership Function Parameters for FIS Input Variables

Input Variable	Linguistic Term	a	b	c
Textual Risk	Low	0.00	0.00	0.50
Textual Risk	Medium	0.00	0.50	1.00
Textual Risk	High	0.50	1.00	1.00

Emotion	Low	0.00	0.00	0.50
Emotion	Medium	0.00	0.50	1.00
Emotion	High	0.50	1.00	1.00
Semantic Mismatch	Low	0.00	0.00	0.50
Semantic Mismatch	Medium	0.00	0.50	1.00
Semantic Mismatch	High	0.50	1.00	1.00
Deepfake Probability	Low	0.00	0.00	0.50
Deepfake Probability	Medium	0.00	0.50	1.00
Deepfake Probability	High	0.50	1.00	1.00

3.6.3 Output Membership Functions

The FIS produces a continuous integrity-risk variable R_f over the normalized interval $[0, 1]$. Four linguistic output states are used:

- Highly Authentic
- Moderately Authentic
- Moderate Fake Risk
- High Fake Risk

Their membership parameters are presented in Table 2.

Table 2. Output Membership Function Parameters

Output Linguistic Term	a	b	c
Highly Authentic	0.00	0.00	0.30
Moderately Authentic	0.20	0.40	0.60
Moderate Fake Risk	0.40	0.60	0.80
High Fake Risk	0.70	1.00	1.00

The overlapping membership functions allow adjacent risk states to be represented gradually rather than as mutually exclusive intervals.

3.6.4 Reduced Expert-Defined Rule Base

The rule base is constructed to capture the principal relationships among multimodal evidence variables. In general, consistent high-risk evidence across multiple modalities produces a stronger risk response, whereas isolated or conflicting evidence can result in an intermediate risk state.

Representative rules from the reduced rule base are given below.

Rule 1: IF Textual Risk is High AND Semantic Mismatch is High AND Deepfake Probability is High THEN Integrity Risk is High Fake Risk.

Rule 2: IF Textual Risk is High AND Emotion is High AND Semantic Mismatch is High THEN Integrity Risk is High Fake Risk.

Rule 3: IF Textual Risk is Medium AND Emotion is Medium AND Semantic Mismatch is Medium AND Deepfake Probability is Medium THEN Integrity Risk is Moderate Fake Risk.

Rule 4: IF Textual Risk is Low AND Emotion is Low AND Semantic Mismatch is

Low AND Deepfake Probability is Low THEN Integrity Risk is Highly Authentic.

Rule 5: IF Textual Risk is Low AND Semantic Mismatch is Medium AND Deepfake Probability is Low THEN Integrity Risk is Moderately Authentic.

Rule 6: IF Textual Risk is High AND Semantic Mismatch is Low AND Deepfake Probability is Low THEN Integrity Risk is Moderate Fake Risk.

Rule 7: IF Textual Risk is Medium AND Semantic Mismatch is High AND Deepfake Probability is Medium THEN Integrity Risk is High Fake Risk.

Rule 8: IF Textual Risk is Low AND Semantic Mismatch is High AND Deepfake Probability is High THEN Integrity Risk is High Fake Risk.

These rules represent representative patterns rather than the complete theoretical 81 combinations. Rules involving combinations that are redundant or contribute marginally to the intended risk interpretation are omitted from the reduced expert-defined rule base.

3.6.5 Fuzzy Inference and Aggregation

For each rule k , a weight-adjusted firing strength is calculated by applying the Mamdani minimum operator to the weighted membership degrees of its antecedents. For a rule involving the four evidence variables, the firing strength is given by

$$\alpha_k = \min \left(w_1 \mu_{C_{text}}^{(k)}, w_2 \mu_E^{(k)}, w_3 \mu_{C_{mismatch}}^{(k)}, w_4 \mu_{P_{fake}}^{(k)} \right)$$

where $\mu^{(k)}$ denotes the membership degree of the corresponding input variable in the linguistic state specified by rule k , and w_i denotes the associated evidence weight. For reduced rules containing fewer antecedents, only the explicitly specified weighted membership terms are included in the firing-strength calculation.

The firing strength is then used to truncate the membership function of the corresponding output linguistic category. For rule k , the truncated consequent is expressed as

$$\mu_k(r) = \min(\alpha_k, \mu_{B_k}(r))$$

where $\mu_{B_k}(r)$ represents the output membership function associated with the consequent of rule k , and $r \in [0, 1]$ denotes the continuous integrity-risk domain. The truncated consequents of all activated rules are subsequently aggregated using the maximum operator:

$$\mu_R(r) = \max_k \mu_k(r)$$

where $\mu_R(r)$ represents the aggregated fuzzy integrity-risk membership function.

Finally, centroid defuzzification is applied to obtain the continuous integrity-risk score:

$$R_{final} = \frac{\int_0^1 r \mu_R(r) dr}{\int_0^1 \mu_R(r) dr}.$$

The resulting R_{final} represents the continuous integrity-risk estimate generated by the weight-adjusted fuzzy reasoning layer. This formulation retains partial membership across linguistic states while incorporating the relative contributions assigned to the four evidence sources during fuzzy rule activation.

3.6.6 Integrity-Risk Interpretation

The continuous risk score R_{final} obtained through centroid defuzzification is mapped to four interpretable integrity-risk categories, as shown in Table 3.

Table 3. Integrity-Risk Interpretation

Risk Score Range	Interpretation
0.00–0.30	Highly Authentic
0.30–0.50	Moderately Authentic
0.50–0.70	Moderate Fake Risk
0.70–1.00	High Fake Risk

The fuzzy formulation permits overlapping evidence states to be represented without imposing an abrupt binary boundary. For example, when textual evidence and semantic mismatch indicate moderate risk while the visual manipulation probability remains low, the resulting assessment can remain in an intermediate risk region rather than being forced directly into an extreme category. The relative contributions of the four evidence

variables are incorporated directly into fuzzy rule activation through their assigned weights.

The Mamdani inference mechanism then integrates the weighted evidence according to the reduced expert-defined rule base, allowing consistent, intermediate, and conflicting evidence patterns to contribute to the final assessment. The resulting continuous R_{final} is used for subsequent risk categorization and quantitative evaluation described in the Results section.

3.7 Multimodal Digital Media Integrity Risk Assessment Pipeline

The overall workflow of the proposed framework is summarized in Algorithm 1, which describes the stepwise process for evaluating the integrity risk of multimodal digital media. The pipeline integrates textual, emotional, OCR-derived, visual, and image–text semantic evidence and combines these heterogeneous evidence sources through a weight-adjusted Mamdani fuzzy inference mechanism. The framework produces a continuous integrity-risk score, an associated confidence measure, a graded risk interpretation, and supporting explainability outputs.

Algorithm 1: Multimodal Digital Media Integrity Risk Assessment Pipeline

Input:

Multimodal media sample $M = \{T, I\}$, where T denotes textual content and I denotes image content.

Output:

Final integrity-risk score R_{final} , confidence measure $Conf$, risk interpretation, and explainability outputs.

Step 1: Input Acquisition

1.1 Receive the textual and image components of the media sample M .

1.2 Identify the available evidence required by the corresponding analytical modules.

Step 2: Textual Analysis

2.1 Process the textual content using the dual-path BERT-based textual analysis module.

2.2 Obtain the stylistic manipulation probability P_{style} and factual-claim inconsistency probability P_{claim} .

2.3 Compute the textual risk score C_{text} using the formulation defined in Section 3.2.

Step 3: Emotion Analysis

3.1 Apply the BERT-based multi-label emotion classifier to obtain the emotion probability vector.

3.2 Apply the neutral suppression coefficient $\lambda = 0.3$.

3.3 Compute the emotion risk score E_{score} according to the formulation defined in Section 3.2.

Step 4: OCR-Based Text Extraction

4.1 Apply Tesseract OCR to the image I to extract embedded text.

4.2 Preprocess the extracted text to remove noise and irrelevant characters.

4.3 Discard OCR text fragments containing fewer than five characters.

4.4 Append the retained OCR text to the associated textual input, as defined in Section 3.3.

4.5 Where OCR-derived text is incorporated into the textual representation, update the relevant textual analysis outputs accordingly.

Step 5: Image–Text Semantic Verification

5.1 Generate text and image embeddings using CLIP.

5.2 Compute cosine similarity between the two embeddings.

5.3 Apply the normalized similarity formulation defined in Section 3.4.

5.4 Obtain the image–text semantic mismatch evidence $C_{mismatch}$.

Step 6: Visual Integrity Assessment

6.1 Process the image using the fine-tuned ResNet50 classifier.

6.2 Obtain the real and manipulated class logits.

6.3 Compute the manipulated-image probability P_{fake} using the softmax formulation defined in Section 3.5.

6.4 Generate the Grad-CAM visualization for interpretation of the visual prediction.

Step 7: Weight-Adjusted Multimodal Fuzzy Fusion

7.1 Assign the fusion weights

$$w_1 = 0.40, w_2 = 0.10, w_3 = 0.25, w_4 = 0.25$$

corresponding to textual risk, emotion risk, semantic mismatch, and visual manipulation probability, respectively.

7.2 Verify that

$$w_1 + w_2 + w_3 + w_4 = 1.$$

7.3 Use C_{text} , E_{score} , $C_{mismatch}$, and P_{fake} as the four normalized inputs to the Mamdani FIS.

7.4 Fuzzify the four evidence variables using the Low, Medium, and High linguistic terms and the membership functions specified in Section 3.6.

7.5 Recognize the $3^4 = 81$ theoretical linguistic combinations.

7.6 Apply the reduced expert-defined rule base specified in Section 3.6.

7.7 Calculate the weight-adjusted firing strength of each applicable rule using the weighted membership degrees:

$$\alpha_k = \min \left(w_1 \mu_{C_{text}}^{(k)}, w_2 \mu_{E_{score}}^{(k)}, w_3 \mu_{C_{mismatch}}^{(k)}, w_4 \mu_{P_{fake}}^{(k)} \right).$$

For reduced rules containing fewer antecedents, only the explicitly specified weighted membership terms are included.

7.8 Truncate the corresponding consequent membership functions according to the rule firing strengths.

7.9 Aggregate the activated consequents using the maximum operator.

7.10 Apply centroid defuzzification to obtain the continuous final integrity-risk score R_{final} .

7.11 Obtain the fuzzy-output-based confidence measure $Conf$, where defined by the implemented inference procedure.

Step 8: Explainability Generation

8.1 Apply SHAP to identify influential textual features associated with the textual-risk prediction.

8.2 Use the Grad-CAM output generated in Step 6 to identify influential visual regions.

8.3 Associate the textual and visual explanations with the corresponding multimodal evidence to form the explanation output.

Step 9: Risk Interpretation

9.1 If $0.00 \leq R_{final} < 0.30$, assign **Highly Authentic**.

9.2 If $0.30 \leq R_{final} < 0.50$, assign **Moderately Authentic**.

9.3 If $0.50 \leq R_{final} < 0.70$, assign **Moderate Fake Risk**.

9.4 If $0.70 \leq R_{final} \leq 1.00$, assign **High Fake Risk**.

Step 10: Output

10.1 Return the final integrity-risk score R_{final} .

10.2 Return the associated confidence measure $Conf$.

10.3 Return the corresponding risk interpretation category.

10.4 Return SHAP-based textual explanations and Grad-CAM visualizations.

10.5 Use the resulting outputs for evaluation on the designated benchmark and external validation datasets.

3.8 Experimental Setup

The proposed framework was evaluated using benchmark datasets, an independent external dataset, repeated experimental runs, statistical analysis, sensitivity analysis, and comparative experiments. The evaluation protocol was designed to examine predictive performance, robustness, generalization, and the contribution of the multimodal risk-fusion mechanism. Calibration-specific performance was not evaluated because dedicated calibration experiments were not conducted; therefore, metrics such as Brier Score and Expected Calibration Error (ECE) are not reported.

3.8.1 Dataset Description and Partitioning

The experimental study employed publicly available datasets representing textual, multimodal, and visual misinformation. The LIAR dataset contains 12,836 annotated political statements categorized into six truthfulness classes (Wang, 2017). FakeNewsNet (Shu et al., 2020) was used to provide news content and associated contextual information, while Fakeddit was

used for multimodal text–image misinformation analysis (Nakamura et al., 2020). For visual integrity assessment, approximately 24,000 images representing authentic, AI-generated, and manipulated content were considered.

To assess generalization beyond the benchmark data, an independent external dataset comprising 1,500 recently collected multimodal samples obtained from public fact-checking sources was used exclusively for external validation. These samples were kept separate from all model-development procedures.

The benchmark data were partitioned using stratified sampling into 70% training, 15% validation, and 15% test sets, preserving the class distribution across the partitions. The validation set was used for model selection and configuration decisions, including fusion-parameter selection and fuzzy-inference-system configuration. The test set remained unseen during model development and was used only for final performance evaluation. The external dataset was not used for training, parameter tuning, or model selection.

3.8.2 Baseline Models

The proposed framework was compared with representative unimodal and multimodal misinformation-detection approaches, including SVM, BiLSTM, BERT-based models, MVAE/MVA, SpotFake+, multimodal transformer models, and CLIP-based approaches. These baselines were selected to provide comparisons across conventional machine-learning, neural, transformer-based, and multimodal modelling paradigms (Singhal et al., 2019).

Recent vision–language and multimodal approaches were additionally considered for literature-level positioning, subject to the availability of compatible implementations and sufficiently comparable evaluation settings. In accordance with this constraint, numerical comparisons with independently reported results were not treated as direct head-to-head comparisons when datasets, partitions, preprocessing, or evaluation protocols differed.

For experimentally evaluated baselines, the same dataset partitions and evaluation

protocol were applied wherever technically applicable. This ensured that differences in measured performance could be examined under a common experimental setting rather than being attributed solely to differences between independently reported studies.

3.8.3 Model Configuration

The principal training and inference configurations used in the experiments are summarized in Table 4. Hyperparameters for trainable models were selected using the validation partition. AdamW was used for the BERT-based models, while the visual classifier was optimized using cross-entropy loss. The pretrained CLIP ViT-B/32 model was used without additional fine-tuning.

Table 4. Model Configuration Details

Module	Configuration
BERT Textual Model	Learning rate = 2×10^{-5} , batch size = 16, epochs = 5, optimizer = AdamW
Emotion Classifier	BERT-based multi-label classifier with sigmoid activation
CLIP Module	Pretrained ViT-B/32 model; used for image–text semantic verification without fine-tuning
Deepfake Detector	ResNet50; learning rate = 1×10^{-4} , batch size = 32, epochs = 20, loss = cross-entropy
Fuzzy Inference Layer	Mamdani FIS with triangular membership functions and a reduced expert-defined rule base
Defuzzification	Continuous centroid method

The configuration of the fuzzy inference layer follows the formulation described in Section 3.6. The four evidence variables are processed using the predefined linguistic representation and reduced rule base before producing the final continuous risk score. The corresponding fusion weights are 0.40, 0.10, 0.25, and 0.25, respectively, and sum to one. These values are kept fixed across the reported experiments.

3.8.4 Implementation and Hardware

The framework was implemented in Python using PyTorch and TensorFlow. Model training and inference were conducted on a CUDA-enabled NVIDIA GPU. The same preprocessing pipeline, model configurations, dataset partitions, and evaluation procedures were maintained across the comparative experiments wherever applicable.

To examine the effect of experimental variability, the framework was evaluated using different random seeds across repeated runs, as described in Section 3.8.6. This

procedure provides a basis for reporting the mean performance and variability of the principal evaluation metrics.

3.8.5 Evaluation Protocol

The predictive performance of the evaluated models was measured using Accuracy, Precision, Recall, and F1-score. These metrics provide complementary measures of overall classification performance and class-sensitive prediction behaviour. Confusion matrices were additionally examined to identify patterns of correct and incorrect predictions across the evaluated classes.

Because the proposed framework produces a continuous integrity-risk score, evaluation was not restricted to the final class label. The analysis also examines the resulting risk states to assess how the framework responds to different combinations of multimodal evidence and whether intermediate risk conditions are preserved rather than reduced to a strictly binary decision.

In addition to threshold-dependent classification metrics, Receiver Operating Characteristic (ROC) analysis was performed using the continuous final integrity-risk score. The Area Under the ROC Curve (ROC-AUC) was used to quantify threshold-independent discrimination between the evaluated classes. The proposed framework achieved a ROC-AUC of 0.97 on the evaluated benchmark test set. This analysis complements the Accuracy and F1-score by assessing the model's discrimination behaviour across different decision thresholds.

Performance was evaluated on both the held-out benchmark test set and the independent external validation dataset. The benchmark test set was used for the principal comparative evaluation, whereas the external dataset was reserved for assessing generalization to previously unseen multimodal samples. The ROC-AUC value reported above corresponds to the benchmark evaluation and should not be interpreted as evidence of equivalent threshold-independent performance on unseen external data.

3.8.6 Statistical Validation

Each principal experiment was conducted over five independent runs using different random seeds. The mean and standard deviation were calculated for the primary

evaluation metrics, and 95% confidence intervals were estimated from the repeated-run results.

Statistical comparisons were performed between the proposed framework and selected baseline models using the same evaluation data and repeated-run protocol. A two-sided significance level of $p < 0.05$ was adopted to determine statistical significance.

The statistical analysis was intended to determine whether the observed differences in performance were consistent across repeated experiments rather than arising solely from run-to-run variability. The resulting statistical comparisons are reported in Section 4.4.

3.8.7 Sensitivity Analysis

A sensitivity analysis was conducted to examine the effect of varying the relative contributions of the four evidence sources within the weight-adjusted fuzzy inference mechanism. The fusion weights were varied while maintaining the normalization constraint

$$w_1 + w_2 + w_3 + w_4 = 1.$$

The evaluated weighting configurations and their corresponding F1-scores are presented in Section 4.5. The analysis assesses the stability of the proposed fusion mechanism and examines the extent to which moderate changes in the relative evidence weights influence the overall predictive performance.

The principal configuration, with weights $w_1 = 0.40$, $w_2 = 0.10$, $w_3 = 0.25$, and $w_4 = 0.25$, was retained for the main experiments based on the sensitivity results. Throughout the analysis, the fuzzy membership functions and reduced expert-defined rule base were kept unchanged, while only the fusion weights were varied. Thus, the observed performance differences reflect changes in the relative weighting of the evidence sources within the fuzzy inference process.

3.8.8 External Validation and Deployment Assessment

Generalization was evaluated using the independent external dataset described in Section 3.8.1. No retraining, parameter adjustment, or model selection was performed

using the external samples. Consequently, the external evaluation provides a separate assessment of the framework under data conditions different from those used during model development.

Performance on the external dataset was compared with the held-out benchmark test results to examine the extent of performance variation under distributional differences. The external results are reported separately to distinguish the primary benchmark evaluation from independent validation.

Practical feasibility was considered through the computational behaviour of the framework, including inference time and generation of explainability outputs. These observations are presented as supplementary implementation characteristics rather than as evidence of deployment readiness.

4. Results and Discussion

The proposed soft-computing-based multimodal framework was evaluated on the benchmark datasets and independent external data described in Section 3.8. The analysis focuses on the effectiveness of the individual evidence-extraction components, the contribution of multimodal fusion, the effect of fusion-weight variation, and the behaviour of the fuzzy risk-assessment mechanism. Repeated-run results are reported where applicable to capture variability across experimental runs.

4.1 Quantitative Performance Evaluation

Performance was evaluated using Accuracy, Precision, Recall, and F1-score. For module-level experiments, the reported values represent the mean $\pm$ standard deviation over five independent runs. The results are presented separately for the textual component and the subsequent multimodal evaluation to distinguish component-level performance from the final integrated framework.

4.1.1 Textual Credibility Module

Table 5 presents the performance of the proposed Dual-BERT textual model in comparison with conventional machine-learning approaches and a single-path BERT model. The comparison examines whether combining linguistic-style and factual-claim

evidence provides additional information for textual misinformation assessment.

Table 5. Performance of Textual Credibility Module (Mean $\pm$ SD)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naïve Bayes	78.4 $\pm$ 1.8	76.9 $\pm$ 1.9	75.8 $\pm$ 2.1	76.3 $\pm$ 1.8
SVM	81.2 $\pm$ 1.5	80.5 $\pm$ 1.7	79.3 $\pm$ 1.6	79.9 $\pm$ 1.5
BERT (Single Path)	88.6 $\pm$ 0.9	87.9 $\pm$ 1.0	88.1 $\pm$ 0.8	88.0 $\pm$ 0.9
Proposed Dual-BERT	92.3 $\pm$ 0.6	91.7 $\pm$ 0.7	92.0 $\pm$ 0.6	91.8 $\pm$ 0.7

The proposed Dual-BERT configuration obtains an F1-score of 91.8%, compared with 88.0% for the single-path BERT model, corresponding to an absolute improvement of 3.8 percentage points. The corresponding accuracy also increases from 88.6% to 92.3%. These results are consistent with the intended role of the dual-path design, in which linguistic-style and factual-claim signals provide complementary evidence for the textual risk assessment.

4.1.2 Deepfake Detection

Table 6 reports the performance of the evaluated visual-integrity models. The results are expressed as mean $\pm$ standard deviation across five independent runs.

Among the evaluated models, ResNet50 records an F1-score of 93.3%, compared with 89.0% for Xception and 86.7% for VGG16. Its standard deviation of 0.8% for F1-score also indicates relatively low variation across the repeated runs. The result supports the use of ResNet50 as the visual evidence-extraction component of the proposed framework. Its output is subsequently represented as the deepfake probability P_{fake} for the risk-fusion stage described in Section 3.6.

Table 6. Performance of Deepfake Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
VGG16	87.2 $\pm$ 1.4	86.5 $\pm$ 1.5	86.9 $\pm$ 1.3	86.7 $\pm$ 1.4
Xception	89.5 $\pm$ 1.1	88.9 $\pm$ 1.2	89.1 $\pm$ 1.1	89.0 $\pm$ 1.1
ResNet50 (Proposed)	93.8 $\pm$ 0.7	93.1 $\pm$ 0.8	93.5 $\pm$ 0.7	93.3 $\pm$ 0.8

4.1.3 Image–Text Consistency Verification

Table 7 presents the results obtained for image–text consistency verification. The comparison considers approaches with progressively stronger cross-modal representation capabilities.

Table 7. Performance of Cross-Modal Consistency Verification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN + Text Embedding	82.6 ± 1.6	81.9 ± 1.5	82.3 ± 1.7	82.1 ± 1.6
Multimodal Transformer	87.4 ± 1.2	86.8 ± 1.3	87.0 ± 1.1	86.9 ± 1.2
CLIP-Based Verification	91.1 ± 0.8	90.5 ± 0.9	90.9 ± 0.8	90.7 ± 0.8

The CLIP-based approach obtains an F1-score of 90.7%, exceeding the corresponding values of the CNN + Text Embedding and Multimodal Transformer approaches in this experiment. This result indicates that the CLIP representation provides useful evidence for assessing semantic correspondence between associated textual and visual content.

Within the proposed framework, this evidence is converted into the semantic-mismatch variable C_{mismatch} using the normalized similarity formulation defined in Section 3.4. The resulting variable is subsequently supplied to the risk-fusion mechanism rather than being treated as an independent final misinformation decision. This allows semantic inconsistency to contribute jointly with textual, emotional, and visual-integrity evidence.

4.2 Ablation Study

An ablation study was conducted to examine the contribution of the principal evidence sources and the fuzzy reasoning stage. One component was removed at a time while the remaining configuration was kept unchanged. The resulting performance is summarized in Table 8.

Table 8. Ablation Study Results

Configuration	Accuracy (%)	F1-score (%)
Without Emotion Module	93.2	92.8
Without CLIP Verification	92.5	92.0
Without Deepfake Detection	91.8	91.3
Without Fuzzy Risk Layer	93.6	93.1
Full Proposed Framework	95.6	95.1

Removing any of the evaluated components results in a reduction in performance relative to the complete framework. Excluding the emotion module reduces the F1-score from 95.1% to 92.8%, while removing CLIP-based verification results in an F1-score of 92.0%. These reductions indicate that affective and cross-modal semantic evidence provide additional information beyond the textual and

visual signals retained in the corresponding ablated configurations.

The removal of the deepfake-detection component produces the largest observed reduction, with the F1-score decreasing to 91.3% and accuracy to 91.8%. This result indicates that visual-integrity evidence makes a substantial contribution to the integrated assessment. Removing the fuzzy risk layer results in an F1-score of 93.1%, showing that the performance of the complete system is not solely attributable to the individual evidence-extraction modules. The complete framework achieves 95.6% accuracy and 95.1% F1-score. Taken together, the ablation results indicate that the evaluated evidence sources provide complementary information and that retaining the fuzzy reasoning stage contributes to the integrated risk-assessment process. Importantly, the ablation study evaluates the contribution of components under the adopted experimental configuration; it does not establish causal superiority of any individual module across other datasets or configurations.

4.3 Comparison with Recent Multimodal Approaches

To position the proposed framework within recent multimodal misinformation research, representative approaches spanning conventional multimodal architectures, vision-language interaction, large vision-language models (LVLMs), and explainable multimodal systems were considered. The comparison is presented in Table 9. The literature reflects a progression from early multimodal feature integration toward increasingly sophisticated vision-language modelling. MVA and SpotFake+ established multimodal approaches that jointly exploit textual and visual information (Singhal et al., 2020). More recent studies have expanded this direction through vision-language knowledge interaction and large vision-language models. Li et al. (2024) investigated vision-language knowledge interaction for multimodal disinformation detection, while Tahmasebi et al. (2024) examined large vision-language models for multimodal misinformation detection. MMFakeBench (2025) further addressed evaluation of

LVLMS using mixed-source multimodal (2025) incorporated explainability into misinformation data, while TRUST-VL multimodal misinformation assessment.

Table 9. Comparison with Recent Multimodal Misinformation Detection Approaches

Method / Approach	Year	Evaluation / Study Context	Relevance to the Present Framework
MVA (Khattar et al.)	2019	Reported in the original study	Early multimodal visual-textual misinformation detection
SpotFake+ (Singhal et al.)	2020	Reported in the original study	Multimodal textual and visual feature integration
Towards Multimodal Disinformation Detection by Vision-Language Knowledge Interaction (Li et al.)	2024	Reported in the original study	Vision-language knowledge interaction
Multimodal Misinformation Detection Using Large Vision-Language Models (Tahmasebi et al.)	2024	Reported in the original study	LVLMS-based multimodal misinformation analysis
MMFakeBench (Liu et al.)	2025	Benchmark reported in the original study	Mixed-source multimodal evaluation of LVLMS
TRUST-VL (Yan et al.)	2025	Reported in the original study	Explainable multimodal misinformation assessment
Proposed Framework	2026	Accuracy: 95.6%; F1-score: 95.1%	Multimodal evidence fusion with fuzzy uncertainty reasoning and XAI

The proposed framework follows a different architectural emphasis. Rather than relying exclusively on an end-to-end multimodal classifier, it retains intermediate evidence from textual risk, emotion, image-text semantic mismatch, and visual manipulation probability before integrating these signals through the fuzzy risk-inference layer. This design preserves the distinction between evidence extraction and risk interpretation and permits heterogeneous signals to contribute to a continuous risk assessment.

The framework achieves 95.6% accuracy and 95.1% F1-score on the benchmark test set under the experimental protocol described in Section 3.8. These results characterize performance within the present study and should not be interpreted as directly equivalent to values reported by independent studies.

In particular, the studies listed in Table 9 employ different datasets, partitions, preprocessing procedures, architectures, training strategies, and evaluation protocols. Therefore, their published performance values are presented for research positioning rather than direct numerical ranking. The comparison instead highlights the methodological relationship between the present framework and recent developments in multimodal misinformation detection.

The distinguishing emphasis of the proposed approach is the combination of multimodal evidence processing, fuzzy uncertainty-aware risk interpretation, and complementary explainability mechanisms

within a single assessment pipeline. This formulation is intended to retain graded information when evidence sources provide consistent, uncertain, or conflicting signals, rather than reducing the assessment exclusively to a binary authentic/fake decision.

4.4 Statistical Validation and Significance Analysis

The stability of the proposed framework was examined over five independent experimental runs. The resulting mean performance, standard deviation, and 95% confidence intervals are presented in Table 10.

Table 10. Statistical Validation of the Proposed Framework

Metric	Mean (%)	Standard Deviation	95% Confidence Interval (%)
Accuracy	95.6	0.42	95.2–96.0
Precision	95.0	0.39	94.7–95.3
Recall	95.3	0.37	95.0–95.6
F1-score	95.1	0.35	94.8–95.4

The relatively small standard deviations indicate limited variation across the repeated runs. The corresponding confidence intervals are also narrow, providing an estimate of the variability associated with the reported performance.

For comparative statistical analysis, the proposed framework was evaluated against the selected multimodal baselines using the same test data and repeated-run protocol. The results are summarized in Table 11.

Table 11. Performance Across Repeated Runs

Model	Accuracy (%)	F1-score (%)
Proposed Framework	95.6 ± 0.4	95.1 ± 0.5
CLIP-guided Model	93.1 ± 0.6	92.7 ± 0.7
Multimodal Transformer	92.4 ± 0.8	91.8 ± 0.9

The proposed framework obtained mean accuracy and F1-score of 95.6% and 95.1%, respectively, across the five runs. The corresponding values for the CLIP-guided model and multimodal transformer were 93.1% and 92.7%, and 92.4% and 91.8%, respectively.

Two-sided paired t -tests were subsequently conducted using the repeated-run results. A significance threshold of $p < 0.05$ was adopted, consistent with the statistical protocol described in Section 3.8.6. The resulting comparisons are given in Table 12.

Table 12. Statistical Significance Analysis

Comparison	p-value
Proposed Framework vs. Multimodal Transformer	0.003
Proposed Framework vs. CLIP-guided Model	0.011

Both reported p -values are below the adopted 0.05 threshold. Under the stated

Table 13. Sensitivity Analysis of Fusion Weights

Configuration	Text Weight	Emotion Weight	Semantic Mismatch Weight	Deepfake Weight	F1-score (%)
C1	0.50	0.10	0.20	0.20	93.8
C2	0.45	0.10	0.225	0.225	94.6
C3 (Proposed)	0.40	0.10	0.25	0.25	95.1
C4	0.35	0.10	0.275	0.275	94.7
C5	0.30	0.10	0.35	0.25	94.1

The adopted configuration, C3, assigns weights of 0.40 to textual risk, 0.10 to emotion risk, 0.25 to semantic mismatch, and 0.25 to deepfake probability, and achieves an F1-score of 95.1%. The alternative configurations produce F1-scores ranging from 93.8% to 94.7%. These results demonstrate that changes in the relative weighting of the evidence sources influence the predictive performance of the framework.

When the textual-risk weight is increased to 0.50 in C1, the F1-score decreases to 93.8%, while the intermediate configuration C2, with a textual weight of 0.45, achieves 94.6%. Reducing the textual weight to 0.35 in C4 results in an F1-score of 94.7%, whereas the configuration C5, with a textual weight of 0.30 and increased semantic-mismatch contribution, achieves 94.1%. Thus, within the configurations evaluated, the observed performance varies with the distribution of weights across the evidence sources.

The results indicate that the evaluated performance is influenced by the balance

testing procedure, the observed differences between the proposed framework and the two selected comparison models are statistically significant.

4.5 Sensitivity Analysis of Risk Fusion Weights

The sensitivity analysis examined the effect of varying the relative contributions of the four evidence sources within the weight-adjusted Mamdani fuzzy inference mechanism. The analysis was conducted under the same evaluation conditions while maintaining the unit-sum constraint on the fusion weights. The fuzzy membership functions and reduced expert-defined rule base were kept unchanged across all configurations. The results are presented in Table 13.

among the four evidence sources rather than by textual evidence alone. In the adopted configuration, textual risk receives the largest individual weight, while semantic mismatch and deepfake probability receive equal weights and emotion risk receives a smaller weight. Based on the tested configurations, C3 was retained for the principal experiments. The sensitivity analysis is intended to assess weight stability within the evaluated configurations. It does not establish that the selected weights are globally optimal or that the same weighting would necessarily produce comparable performance across different datasets, domains, or operating conditions.

4.6 External Validation and Real-World Applicability

To examine generalization beyond the benchmark evaluation, the proposed framework was tested on approximately 1,500 previously unseen multimodal samples collected from publicly available fact-

checking and media-verification sources. The external samples covered multiple content domains, including politics, public health, social issues, and current events. None of these samples was used during training, parameter selection, or model configuration.

The results are presented in Table 14.

Table 14. External Validation Results

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Benchmark Test Set	95.6	95.0	95.3	95.1
External Validation Dataset	93.4	92.8	93.1	92.9

Performance on the external dataset decreases relative to the benchmark test set, with F1-score changing from 95.1% to 92.9%. The reduction is consistent with the expected effect of evaluating previously unseen data from a different collection context. At the same time, the external results remain within the reported performance range of the framework, providing evidence of generalization under the evaluated external conditions.

Qualitative examination of the external samples also indicated that the framework could identify cases involving visual manipulation and inconsistencies between textual and visual content. These observations are qualitative and are therefore not used as additional quantitative evidence of performance. The measured average inference time was approximately 0.42 seconds per multimodal sample in the GPU-enabled experimental environment. This value represents a measurement under the specific hardware and workload used in the study and should not be interpreted as a general real-time deployment guarantee. Evaluation across different hardware configurations, workloads, and larger-scale operational settings would be required before making broader deployment claims.

4.7 ROC Analysis

The ROC analysis provides a threshold-independent evaluation of the continuous integrity-risk score produced by the proposed framework. Figure 2 illustrates the relationship between the True Positive Rate

(TPR) and False Positive Rate (FPR) across a range of decision thresholds.

The reported ROC analysis yields an Area Under the ROC Curve (AUC) of 0.97, indicating a high level of discrimination between the evaluated classes across varying decision thresholds. Since an AUC of 0.50 corresponds approximately to random ranking, the substantially higher AUC obtained by the proposed framework demonstrates that the generated integrity-risk score provides effective class separation over a broad range of operating thresholds. The ROC result therefore provides evidence of strong threshold-independent discriminative capability within the evaluated test setting.

The ROC analysis is also useful for examining how changes in the decision threshold influence the trade-off between false-positive and true-positive rates. In an operational decision-support setting, threshold selection may depend on the intended application and on the relative consequences associated with false-positive and false-negative decisions. Although the obtained AUC indicates strong overall discrimination, application-specific threshold optimization is beyond the scope of the present study.

The present work does not report Brier Score or Expected Calibration Error (ECE) because dedicated probability-calibration experiments and corresponding numerical results were not retained in the final evaluation. Accordingly, no quantitative claim regarding probability calibration is made. The reported AUC should therefore be interpreted as a measure of discrimination/ranking performance, rather than as evidence of calibrated probability estimates.

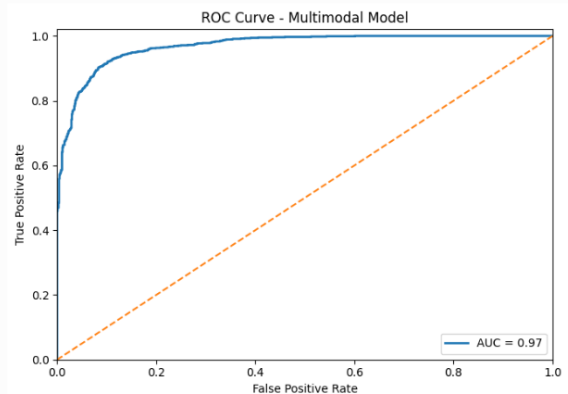


Figure2. ROC Curve Analysis

Overall, the ROC analysis provides an important complementary perspective on the behaviour of the continuous integrity-risk score. The AUC of 0.97 demonstrates strong threshold-independent discrimination in the evaluated test setting, while the benchmark Accuracy, Precision, Recall, and F1-score represent performance at the selected operating point. Together, these results provide complementary evidence regarding the discriminative and operating-point performance of the proposed framework.

4.8 Qualitative Analysis and Explainability

The qualitative analysis examines representative explanation outputs in relation to the corresponding multimodal integrity-risk assessments. The explanation mechanisms provide localized evidence for inspecting the factors associated with the automated assessment: textual attribution highlights influential portions of the input text, whereas visual attribution identifies image regions contributing to the visual-integrity assessment.

These complementary explanations are particularly informative when evidence from different modalities is consistent or conflicting. For instance, elevated textual risk may coexist with comparatively low visual manipulation evidence. Examining the modality-specific explanations together with the final risk assessment allows such discrepancies to be inspected explicitly rather than being obscured by a single categorical output.

The qualitative results therefore demonstrate the interpretive value of the framework for human-in-the-loop verification. The explanations can assist reviewers in identifying which textual or visual evidence contributed to an assessment and in examining cases associated with intermediate risk levels. They are, however, treated as supporting interpretive evidence rather than definitive proof of misinformation or media manipulation. Consequently, the automated risk score and its explanations should be considered together with contextual information and appropriate human verification, particularly in applications

where erroneous assessments may have significant consequences.

This analysis complements the quantitative evaluation by illustrating how the framework exposes modality-specific evidence underlying its risk assessments, without treating the explanation outputs as independent predictions.

4.9 Risk Assessment and Soft Computing Analysis

The proposed Mamdani fuzzy inference layer provides the final soft-computing mechanism for integrating four heterogeneous evidence variables: textual risk, emotion risk, image-text semantic mismatch, and visual manipulation risk. Unlike a discrete decision mechanism in which each input is assigned exclusively to a single category, the fuzzy layer permits partial membership across neighbouring linguistic states. This enables gradual transitions between different levels of assessed integrity risk.

The four-input formulation theoretically produces $3^4 = 81$ combinations of the three linguistic states assigned to each input. Rather than enumerating all possible combinations, the framework employs a reduced expert-defined rule base containing practically meaningful multimodal risk patterns, as specified in Section 3.6. This approach limits redundant rules while retaining combinations considered relevant to the interpretation of heterogeneous integrity-risk evidence.

For each input instance, the fuzzy inference layer determines the membership degree of the four evidence variables in their respective linguistic categories. The corresponding fusion weights are incorporated directly into the rule-activation process. Thus, the relative contributions of textual risk, emotion risk, semantic mismatch, and visual manipulation risk influence the firing strength of the applicable fuzzy rules. The activated consequents are subsequently truncated using the Mamdani minimum operator and aggregated using the maximum operator. Centroid defuzzification then produces a continuous integrity-risk score.

This weight-adjusted inference mechanism allows the contribution of each evidence

source to be considered during fuzzy reasoning rather than treating all evidence variables as having identical influence. At the same time, partial membership is retained,

allowing heterogeneous evidence to contribute to intermediate risk states rather than forcing an immediate binary authentic/fake decision.

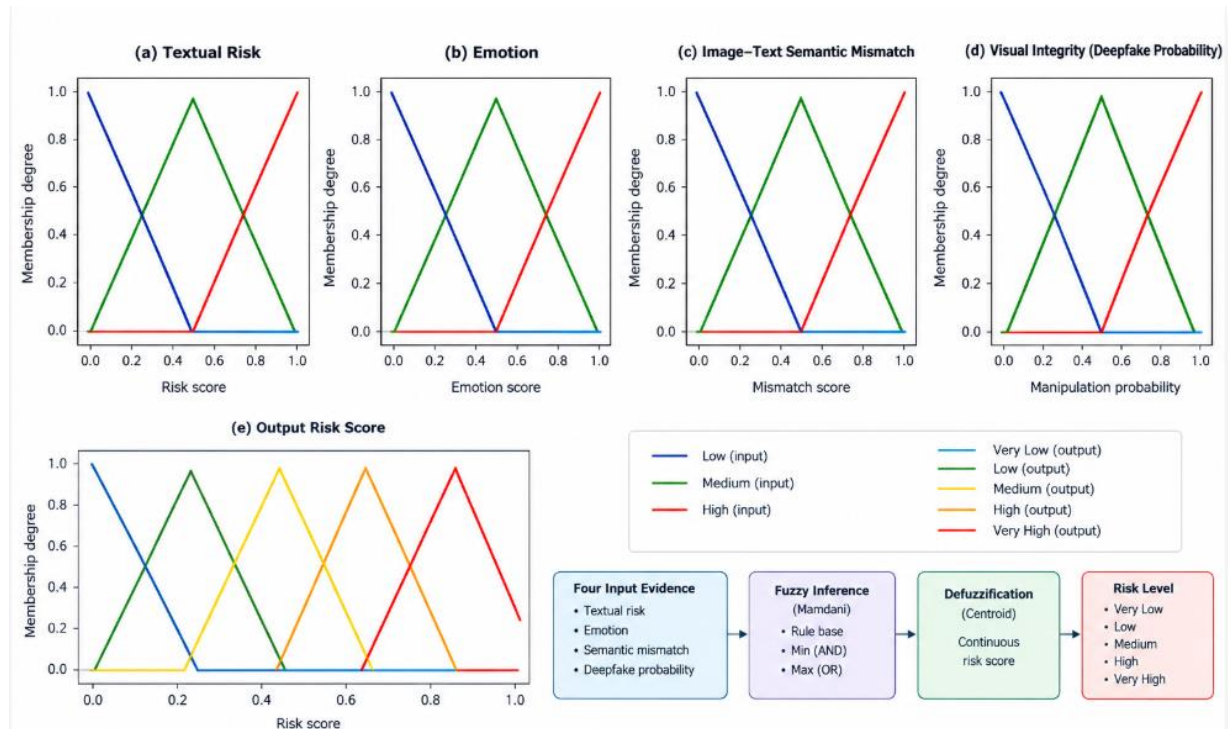


Figure 3. Fuzzy membership functions in the proposed framework

The experimental analysis demonstrates the interpretive role of this mechanism for cases involving intermediate or conflicting evidence. For example, an instance exhibiting moderate image–text semantic inconsistency but relatively low visual manipulation probability can contribute to an intermediate risk assessment rather than necessarily producing an extreme risk state. Similarly, differing evidence levels across textual, emotional, semantic, and visual inputs can activate neighbouring linguistic states with different degrees of membership.

Figure 3 illustrates the linguistic membership functions used by the fuzzy inference layer. The overlap between neighbouring membership functions permits intermediate input values to possess partial membership in more than one linguistic category. Consequently, changes in multimodal evidence can produce gradual changes in the resulting risk assessment rather than abrupt transitions between discrete categories.

The fuzzy layer therefore provides an uncertainty-aware integration mechanism for

heterogeneous evidence generated by the preceding neural and multimodal modules. Its contribution is primarily integrative and interpretive: it combines weighted evidence through expert-defined fuzzy rules and represents the resulting assessment as a continuous risk value. This is particularly relevant when credibility-related signals from different modalities do not consistently indicate either a low- or high-risk condition.

Overall, the risk-assessment stage complements the neural components by transforming heterogeneous multimodal evidence into a graded integrity-risk representation through weight-adjusted fuzzy reasoning. The resulting score can support subsequent human inspection by indicating the degree and linguistic state of the assessed risk. The fuzzy output and associated explanations should, however, be interpreted as decision-support information rather than as definitive evidence of misinformation or digital manipulation.

5. Discussion

The experimental findings indicate that digital media integrity assessment can incorporate complementary information from multiple evidence sources rather than relying exclusively on a single modality. The module-level results demonstrate measurable performance across textual, emotional, semantic, and visual analysis, while the ablation analysis shows changes in performance when individual components are removed. The full framework achieved an F1-score of 95.1%, compared with 92.8% without the emotion module, 92.0% without CLIP-based semantic verification, 91.3% without deepfake detection, and 93.1% without the fuzzy risk layer. These results provide component-level evidence for the contribution of the integrated configuration under the evaluated experimental conditions. They should not, however, be interpreted as establishing causal superiority across all datasets or application settings.

The comparison with recent multimodal approaches provides contextual positioning of the proposed framework. Direct numerical comparisons should be interpreted cautiously because published studies may employ different datasets, data partitions, preprocessing procedures, evaluation metrics, and model configurations. The contribution of the proposed framework therefore does not depend on introducing BERT, CLIP, or ResNet50 as individually novel architectures. Instead, the framework combines established models within a common integrity-risk assessment pipeline. The integration of heterogeneous evidence with fuzzy reasoning provides a mechanism for representing intermediate and potentially conflicting evidence, while SHAP and Grad-CAM provide modality-specific information for inspecting the associated assessments.

The sensitivity analysis further demonstrates that the resulting performance is affected by the relative weighting assigned to the four evidence sources. The adopted configuration assigns weights of 0.40, 0.10, 0.25, and 0.25 to textual risk, emotion risk, semantic mismatch, and visual manipulation risk, respectively, and produces an F1-score of 95.1% under the evaluated conditions.

Because these weights are incorporated directly into fuzzy rule activation, the sensitivity results reflect changes in the relative contribution of the evidence sources within the fuzzy fusion mechanism. The analysis should therefore be understood as an assessment of the tested weighting configurations rather than as evidence of a universally optimal weighting scheme. The repeated-run evaluation showed relatively small variation in the reported performance measures under the experimental protocol. Statistical testing also indicated significant differences for the evaluated comparisons with the CLIP-guided model and multimodal transformer baseline. These results provide evidence of consistency within the tested experimental setting. Nevertheless, the limited number of repeated runs and dependence on the selected benchmark should be considered when interpreting the statistical findings and their generalizability.

External validation provides additional evidence concerning the effect of data distribution on framework performance. The F1-score decreased from 95.1% on the benchmark test set to 92.9% on the external dataset, corresponding to a reduction of 2.2 percentage points. Since the external samples were not used for training, tuning, or model selection, the observed reduction is consistent with performance variation under a different data distribution. It should not be interpreted as establishing universal real-world generalization. Domain shift, changing manipulation techniques, OCR errors, annotation differences, and disagreement between modalities may contribute to performance variation outside the evaluated datasets.

The ROC analysis provides an additional perspective on the threshold-independent behaviour of the proposed continuous integrity-risk score. The framework achieves 95.6% accuracy and an F1-score of 95.1% at the evaluated decision setting, while the ROC analysis produces an AUC of 0.97. This high AUC indicates strong discrimination between the evaluated classes across a broad range of decision thresholds. Thus, unlike the operating-point metrics, which characterize performance at a selected threshold, the AUC

provides evidence that the framework maintains effective ranking and class-separation capability across varying threshold settings. The distinction between these measures remains important: Accuracy and F1-score describe performance at the selected operating point, whereas AUC summarizes threshold-independent discrimination of the continuous risk score.

Taken together, the findings support the use of the proposed framework as an explainable, uncertainty-aware decision-support approach under the evaluated conditions. Its principal contribution is the structured integration of heterogeneous textual, emotional, semantic, and visual evidence through weight-adjusted fuzzy reasoning, together with modality-specific explanation mechanisms. The AUC of 0.97, together with the operating-point performance, provides complementary evidence of strong discriminative behaviour within the evaluated experimental setting. However, the reduction in external F1-score indicates that performance remains sensitive to changes in data distribution, and broader claims concerning real-world generalization should therefore be made cautiously.

Future evaluation should consider larger, temporally separated, independently annotated, and domain-diverse datasets. Further work should also investigate probability calibration, threshold selection under application-specific costs, robustness to evolving manipulation techniques, and the stability of fusion weights across different domains and data distributions. Such evaluations would provide a more comprehensive assessment of the framework's reliability when deployed under heterogeneous real-world conditions.

6. Conclusion

This study presents an explainable multimodal framework for digital media integrity-risk assessment under uncertainty. The framework integrates textual credibility, emotional signals, OCR-derived information, image-text semantic consistency, and visual manipulation evidence through a weight-adjusted Mamdani fuzzy inference mechanism. The framework achieved 95.6% accuracy and 95.1% F1-score on the

benchmark test set, while external validation achieved an F1-score of 92.9%, indicating performance variation under a different data distribution. The ablation and sensitivity analyses provide evidence of performance changes associated with the integrated multimodal components and tested fusion configurations. The selected weights—0.40 for textual risk, 0.10 for emotion risk, 0.25 for semantic mismatch, and 0.25 for visual manipulation risk—were retained based on the evaluated sensitivity configurations. These results describe the behaviour of the framework under the tested conditions and do not establish the weights as universally optimal. SHAP and Grad-CAM provide supporting modality-specific information for interpreting textual and visual assessments. These explanations are intended to assist inspection of the evidence contributing to an assessment and should not be regarded as definitive proof of misinformation or manipulation. The reported ROC-AUC of 0.97 provides additional evidence of strong threshold-independent discrimination of the continuous integrity-risk score across the evaluated range of decision thresholds. The AUC complements the operating-point Accuracy and F1-score by characterizing the ranking and discrimination behaviour of the framework across varying thresholds. The principal contribution of the study is the structured integration of heterogeneous multimodal evidence, weight-adjusted fuzzy uncertainty reasoning, and explainable decision support within a unified integrity-risk assessment framework. The observed reduction in external F1-score highlights the influence of data distribution on framework performance and indicates the need for further validation across diverse real-world conditions. Future work will focus on larger and temporally diverse datasets, independent probability-calibration analysis, robustness to evolving manipulation techniques, lightweight architectures, more systematic threshold selection, and the incorporation of additional modalities such as video and audio.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or

personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

The data used in this study were obtained from publicly available image and video repositories and are subject to their respective licenses. The processed data and annotations generated during this study are available from the corresponding author upon reasonable request.

References

- Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: A survey of deep learning techniques. *Social Network Analysis and Mining*, 13, 101. <https://doi.org/10.1007/s13278-023-01104-w>.
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054. <https://doi.org/10.18653/v1/2020.acl-main.372>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Jin, R., Fu, R., Wen, Z., Zhang, S., Liu, Y., & Tao, J. (2024). Fake news detection and manipulation reasoning via large vision-language models. *arXiv preprint arXiv:2407.02042*.
- Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. *Proceedings of the 2019 World Wide Web Conference*, 2915–2921. <https://doi.org/10.1145/3308558.3313552>
- Li, Q., Gao, M., Zhang, G., Zhai, W., Chen, J., & Jeon, G. (2024). Towards multimodal disinformation detection by vision-language knowledge interaction. *Information Fusion*, 102, 102037. <https://doi.org/10.1016/j.inffus.2023.102037>
- Liu, X., Li, Z., Li, P., Huang, H., Xia, S., Cui, X., Huang, L., Deng, W., & He, Z. (2025). MMFakeBench: A mixed-source multimodal misinformation detection benchmark for LVLMS. *International Conference on Learning Representations (ICLR 2025)*.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- Mamdami, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1), 1–13.
- Nakamura, K., Levy, S., & Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6149–6157.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748–8763.
- Sagar, A. S. M., Bennamoun, M., Boussaïd, F., Sharif, N., Xu, L., Sahmoud, S., & Kishk, A. (2026). Fact or fake? Assessing the role of deepfake detectors in multimodal misinformation detection. *IEEE Access*, 14, 86992–87011. <https://doi.org/10.1109/ACCESS.2026.3686437>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations

- from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3), 171–188. <https://doi.org/10.1089/big.2020.0062>
- Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., & Kumaraguru, P. (2020). SpotFake+: A multimodal framework for fake news detection via transfer learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10), 13915–13916. <https://doi.org/10.1609/aaai.v34i10.7230>
- Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A multimodal framework for fake news detection. *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 39–47. <https://doi.org/10.1109/BigMM.2019.00-44>
- Tahmasebi, S., Müller-Budack, E., & Ewerth, R. (2024). Multimodal misinformation detection using large vision-language models. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*, 2189–2199. <https://doi.org/10.1145/3627673.3679826>
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Wang, W. Y. (2017). “Liar, liar pants on fire”: A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426. <https://doi.org/10.18653/v1/P17-2067>
- Yan, Z., Qi, P., Hsu, W., & Lee, M.-L. (2025). TRUST-VL: An explainable news assistant for general multimodal misinformation detection. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 5588–5604. <https://doi.org/10.18653/v1/2025.emnlp-main.284>
- Yang, P., Ma, J., Liu, Y., & Liu, M. (2023). Multi-modal transformer for fake news detection. *Mathematical Biosciences and Engineering*, 20(9), 16115–16134.
- Yang, Y., Yang, Y., Ying, Q., Qian, Z., & Zhang, X. (2023). Multi-modal fake news detection on social media via multi-grained information fusion. *arXiv preprint arXiv:2304.00827*.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338–353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)