Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

Article - e006

REMIX-FND: A Multi-Modal Domain-Invariant Framework with Adaptive Evidence Retrieval for Cross-Domain Fake News Detection

Priyadarshan Khadtale[1], Dr. Rajesh Bansode[2]

[1]Thakur College of Engineering and Technology University of Mumbai, Maharashtra, India

[2]HOD, Department of Information Technology, Thakur College of Engineering and Technology, University of Mumbai, Maharashtra, India

Received: 03/03/2026 **Revision Received**: 30/03/2026 **Accepted**: 07/04/2026

ABSTRACT

Misinformation is rampant on social media, with public health and civic harms demonstrated 1,2, whereas applied models face domain shift, latency in content retrieval, and machine-produced language 3–6. REMIX-FND is an open-source multimodal pipeline that fuses textual, imagery, social and temporal graph features with the option of adversarial domain adaptation (DANN) and DIML, a DIML variant that merges gradient reversal domain confusion with MAML inner loop training on the same fused feature space. Inference involves a Monte Carlo dropout-based uncertainty metric for variable depth evidence analysis, Dynamic Source Reliability Graphs with time-weighted edges, and output of a six-detector AI-text score, stance classification, and tiered explanation through a standardized API. For a FakeNewsNet-style held-out test partition (n=3,253) under equal epoch budget allocation between methods, averaged experiments reveal a mean weighted F1 score around 84.9–85.3% in baseline, DANN-enhanced, and DIML-modified settings, with inter-method differences less than expected cross-seed standard deviations; HC3-style AI-text evaluation outputs threshold metrics of 0.55 and AUROC 0.777 (average of odd-even line splits from the same snapshot; see Table 9). The code, dataset, and evaluation scripts are made available to facilitate replicable systems-level investigation into multimodal misinformation detection.

KEYWORDS: Fake News Detection, Multi-Modal Fusion, Domain-Adversarial Training, Meta-Learning, Uncertainty-Guided Retrieval, AI-Generated Text Detection

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

1. INTRODUCTION

1.1 Motivation

Despite achieving good performance within domain for standard split benchmarks, existing architectures experience double digit reductions in F1 due to domain shifts at an outlet level 3–5, which is compounded if the training and testing data differ in both topic, outlet, and temporal domains. Experimental evidence demonstrates that false information travels faster, further, and reaches more people on social media than corrective information 1, with public health organizations considering misinformation as a core component of risk communication on par with other operational priorities 2. Thus, platforms and external moderators require models that work under the conditions of domain shift 3–5, with limited latency and compute constraints that may include retrieval and graph models 4, as well as machine-generated and paraphrased text which bypass brittle lexicon-based checks 6. Existing literature tends to address only one aspect, with little focus on reproducible architectures that integrate multimodal fusion, domain adaptation, uncertainty in retrieval, graph-based source modeling, and AI scoring, among others.

1.2 Problem statement

A unified deployable misinformation scoring framework should incorporate multiple modalities for representation learning, optionally domain adaptation, evidence retrieval under bounded latency, auxiliary data (stance, machine-generated text), and an interface specification such that the splits, seeds, and evaluation metrics can be replicated by other entities.

1.3 Proposed approach

REMIX-FND consists of a three-step pipeline (Fig. 1). Step 1 trains multi-modal encoders optionally with domain adversarial learning (+DANN) alignment and optionally with a DIML criterion, which incorporates inner-loop MAML learning in addition to the fused representation. Step 2 computes evidence depth based on Monte Carlo Dropout uncertainty and propagates time-decayed source reliability via DSRG. Step 3 appends a fixed-weight AI-text detector with six classifiers; a LIAR fine-tuned stance classifier head, and layered explanations by a reference API for an early exit option. The package comes with a

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

FakeNewsNet style dataset for tag-stratified analyses; veracity results are presented based on the entire dataset with multi-seed averaging.

Fig.1 REMIX-FND end-to-end pipeline architecture. Processed layers represented as rectangles: multimodal encoders of text, images, social, temporal graph data are fed into fusion layer followed by optional training of DANN and DIML; during inference, MC dropout applied on the veracity module generates uncertainty which decides on shallow/deep evidence gathering; retrieved paragraphs go through the DSRG component and cross-encoder stance aggregation step; final results include veracity, stance, optionally AI/text score and explanation levels exposed via API endpoints.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

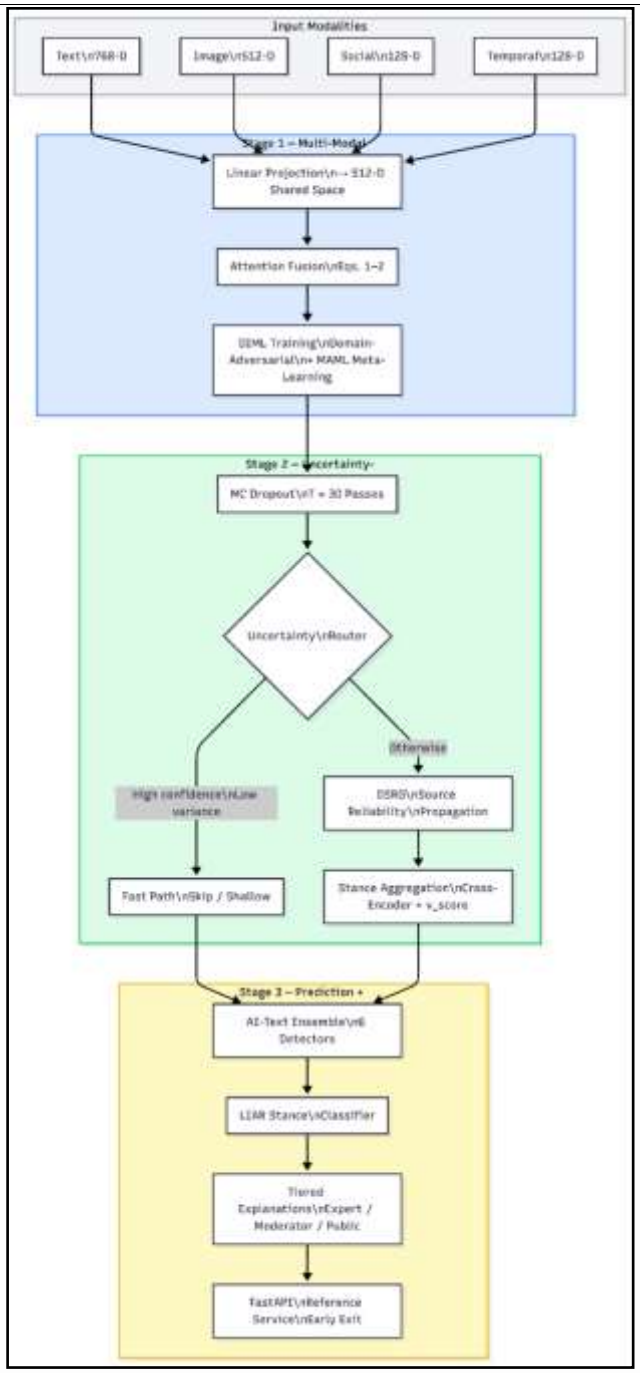

Fig. 1 — Architecture Diagram

1.4 Contributions

Contributions made by this paper include the following. (1) Holistic open system. REMIX-FND is the release of a unified repository integrating four-branch multimodal fusion with learned modality missingness, +DANN, DIML if applicable, uncertainty-conditioned retrieval depth, DSRG, six detector AI-text ensemble, LIAR stance fine-tuning, and the FastAPI

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

reference implementation with corresponding scripts and file splits. (2) Training process dynamics. This paper details the coexistence of adversarial domain confusion with meta-learning inner loop sharing one forward graph, potential interference between +DANN and DIML, and proper epoch budgeting and staged scheduling to prevent an unfair comparison between the two. (3) Empirical benchmarking. The mean ± standard deviation for veracity on the standard test set (n = 3,253) under equivalent epoch caps is provided along with seed numbers per method (given in main table); HC3-style AI-text, macro-F1, and AUROC metrics are provided as well, along with line splits of the JSON data source for AUROC (Table 9).

The rest of the paper discusses relevant literature, methodologies employed, findings observed, and closes with a discussion of limitations and conclusions.

2 LITERATURE REVIEW

Multi-modal approaches to fake news detection leverage multiple streams of information - language, images, social interactions, and temporal relationships—since individual modes can be readily faked or excluded 7–9. Previous research shows the benefits of cross-modal attention and adversarial learning under fixed split conditions; EANN 9 proposes adversarial invariant models for events, whereas UMD² 5 is concerned with weakly supervised multi-modal learning. The proposed approach shares the same spirit but is designed with deployment constraints in mind: four independent encoders, learnable zero vectors for missing features, and optional fine-tuning on a single tensor instead of a selected pair of modalities.

Domain shift is the leading cause of failure when the training and testing distributions differ. Recent work on distributional aware learning, such as DAL 3, has demonstrated the importance of robustness to shifts in topics and outlets. While UMD² 5 and MUSER 4 tackle the challenge from different angles—multi-modal learning and structured evidence—their OOD performance tables still show a 10% drop in F1 scores compared with in-domain performance 3–5. In this paper, we consider the problem of cross-domain testing as a practical challenge—latency, uncertainty handling, and consistent results.

The evidence-based architecture extracts or passes down external textual and graphical context instead of classifying isolated headlines. The MUSER 4 model treats multi-hop evidences as an explicit first-class object, while DAL 3 integrates claim representation learning with adversarial debiasing. REMIX-FND belongs to the same category as MUSER 4

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

and DAL 3 but contrasts shallow and deep search against predictive uncertainty estimation (below) and uses a graph of source reliability to score the extracted evidence.

Uncertainty Quantification (UQ) in NLP ranges from Monte Carlo dropout 12 to post-hoc calibration 13 and ensemble variance, though no method works effectively at predicting true confidence when there is a distributional shift. REMIX-FND uses MC dropout as a simple routing policy based on tractability concerns, not UQ. The model design fits moderate environments where it is unrealistic to run a full ensemble or Bayesian classifier for each instance.

Graph neural networks have emerged as the de facto choice for rumor and misinformation propagation modeling due to their ability to represent the sharing dynamics of the social network—i.e., who shares content when and how. Bi-directional graph convolution-based models distinguish bottom-up and top-down propagation on social media threads, while co-attention models that use comment structures facilitate explainability in detection tasks. Geometry-aware models specifically tackle news-user interaction graphs at time of posting. In contrast, the DSRG model of REMIX-FND is applied to an evidence-source graph constructed for inference purposes and with manually-defined temporal edge decay rather than a social media cascade mined using the Twitter API.

The threat posed by machine-generated text is unique and challenging. Paraphrasing and rewriting attacks severely reduce detector precision without compromising ranking metrics. HC3 and other datasets containing human-written text and ChatGPT responses enable evaluation against machine-generated text using established benchmarking practices 16; however, accuracy scores vary greatly depending on thresholds and similarity of test data to heuristics used to train detectors. Therefore, REMIX-FND reports both AUROC and accuracy at various thresholds for their fixed six-detector ensemble.

LLMs have moved fact-checking from being based on static classifiers to retrieval-enhanced verification and long-form factual scoring. Current efforts measure atomic factual accuracy in generated text using fine-grained scoring pipelines and examine self-consistency for hallucination detection, but industrial-strength analysis focuses on evading detectors by adaptive adversaries 6. REMIX-FND is not meant to replace such pipelines; rather, it reveals an efficient AI-text score and API endpoints to enable systems scientists to plug in better LLM-based verification modules without changing the multimodal backbone.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

Table 1 compares REMIX-FND against representative previous systems on integration dimensions that are critical to replicability and adoption (coverage of modality types, domain adaptation, uncertainty-awareness of retrieval, graph scoring of evidence, auxiliary AI-text scoring, explanations surfaces, and open code). The rows do not represent a merit hierarchy; existing approaches vary by backbone models, split conditions, and evidence budgets. The comparison assertion is limited to the following: REMIX-FND is the sole column that combines all these capabilities within one open repository with a consistent evaluation pipeline.

Table 1 REMIX-FND against previous systems

Feature	EANN 9	MUSER 4	DAL 3	UMD² 5	REMIX-FND
Multi-modal fusion	✓			✓	✓
Domain-adversarial training	✓		✓		✓
Meta-learning (MAML / DIML coupling)					✓
MC dropout uncertainty routing					✓
Graph source reliability		✓			✓
AI-text detection ensemble					✓
Tiered explanations					✓
Open deployable codebase					✓

3 METHODOLOGY

3.1 Architecture, fusion, and DIML

The text (768-D pooling), image (512), social MLP (128), and temporal graph attention (128) are projected into the common 512-D fusion space. Null embeddings are also learned for attention to down-weight the absence of other modalities. For each modality i in text, image, social, and graph with df = 512, and ziPihi is the corresponding stacked feature representation.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

$$\alpha \; softmax(W\text{attn}\; z\; b\text{attn}) \in \mathbb{R}4 \quad (1)$$

$$h\text{fused} \; \Sigma\text{i} \; \alpha\text{i} \; z\text{i} \in \mathbb{R}\text{df}. \quad (2)$$

In the domain adversarial approach 11, let G denote the feature extractor and Ddomain the domain classifier. Denote the usual domain classification loss as Ld,

$$L\text{d} - E(x,d)\log P_D(\text{dG(x)})^*) \quad (3)$$

The gradients w.r.t. G traverse the gradient reversal layer (GRL), such that G is optimized to reduce the prediction of d by the domain classifier, or equivalently optimize for domain-invariance features in the sense of the min-max problem of Ganin et al. We represent the component in Ltotal as Ladv ≡ Ld from Eq. (3).

For meta-learning, we follow MAML 10 in which outer optimization is conducted for θ, and for each domain d, K 5 iterations (as set as default in Finn et al. 10; K is not ablated in this study) are performed using α 5×10−6 over the support set Sd. The inner-loop recurrence is

$$\theta\text{d(k+1)} \; \theta\text{d(k)} - \alpha \; \nabla\theta L\text{support}(f\theta\text{d(k)}; \; S\text{d}), \; \text{k } 0, \ldots, \text{K}-1^*) \quad (4)$$

With $\theta'\text{d} := \theta\text{d(K)}$,

$$L\text{meta} \; E\text{d}L\text{query}(f\theta'\text{d}; \; Q\text{d}) \quad (5)$$

Lquery is the same veracity (classification) loss from the primary task but now applied to the query set Qd. Total loss for the multi-objective approach becomes

$$L\text{total} \; L\text{cls} \; \lambda\text{adv}L\text{adv} \; \lambda\text{meta}L\text{meta}, \quad (6)$$

Novelty in architecture relative to “DANN + MAML” stacks. Previous methods for domain generalization generally employ meta-learning objectives for learning cross-domain classifiers, but do not deploy gradient-reversal-based domain confusion on the same shared representation or combine a pre-trained DANN encoder with a meta-learner in a pipeline alternation fashion. The novel contribution of the DIML used in REMIX-FND is its joint deployment of adversarial debiasing and fast adaptation on the same multi-modal fusion encoder in one optimization cycle (see Algorithm 1). This architecture involves the interplay of three objectives using one common backbone: (i) veracity classification Lcls, (ii) domain confusion Ladv based on GRL on the very same tensors that serve the veracity classifier, and (iii) MAML-inspired support loss Lsupport and query loss Lmeta on top of them that affect shared weights. The architecture is responsible for the coupling of the two mechanisms: the

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

gradients from the inner loop pass through the GRL-debiased trunk so that fast adaptation and domain invariance interact simultaneously—not as a result of sequential composition of two algorithms or independent training cycles (Fig. 2).

Algorithm 1: One DIML Outer Update

Require: Encoder G, veracity head C, domain classifier D,
GRL with reversal strength λ_rev,
inner step size α, outer step size β,
hyperparameters λ_adv, λ_meta, K inner steps

1. Sample mini-batch B with labels y_B, domain labels d_B
2. Sample per-domain support/query pairs {(S_d, Q_d)}
// Batch-level losses — GRL active
3. L_cls ← CrossEntropy(C(G(B)), y_B)
4. L_adv ← CrossEntropy(D(GRL(G(B))), d_B)
// Inner adaptation per domain
5. For each domain d:
a. Set θ_d^(0) ← θ
b. For k = 0, …, K−1:
L_sup ← CrossEntropy(C(G_{θ_d^(k)}(S_d)), y_Sd)
+ λ_adv · CrossEntropy(D(GRL(G_{θ_d^(k)}(S_d))), d_Sd)
// GRL is active during inner steps:
// fast adaptation and domain confusion are jointly constrained
θ_d^(k+1) ← θ_d^(k) − α ∇_{θ_d^(k)} L_sup
c. Set θ'_d ← θ_d^(K)
// Meta loss on adapted parameters
6. L_meta ← E_d[CrossEntropy(C(G_{θ'_d}(Q_d)), y_Qd)]
// Combined outer update
7. L_total ← L_cls + λ_adv · L_adv + λ_meta · L_meta
8. θ ← θ − β ∇_θ L_total

Since the GRL is active throughout the inner steps (Line 5b), the inner adaptation minimizes task losses and enhances domain confusion on the support set, and this combined constraint makes the proposed method distinct from other methods using an alternating process between adversarial and meta-losses in two stages.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

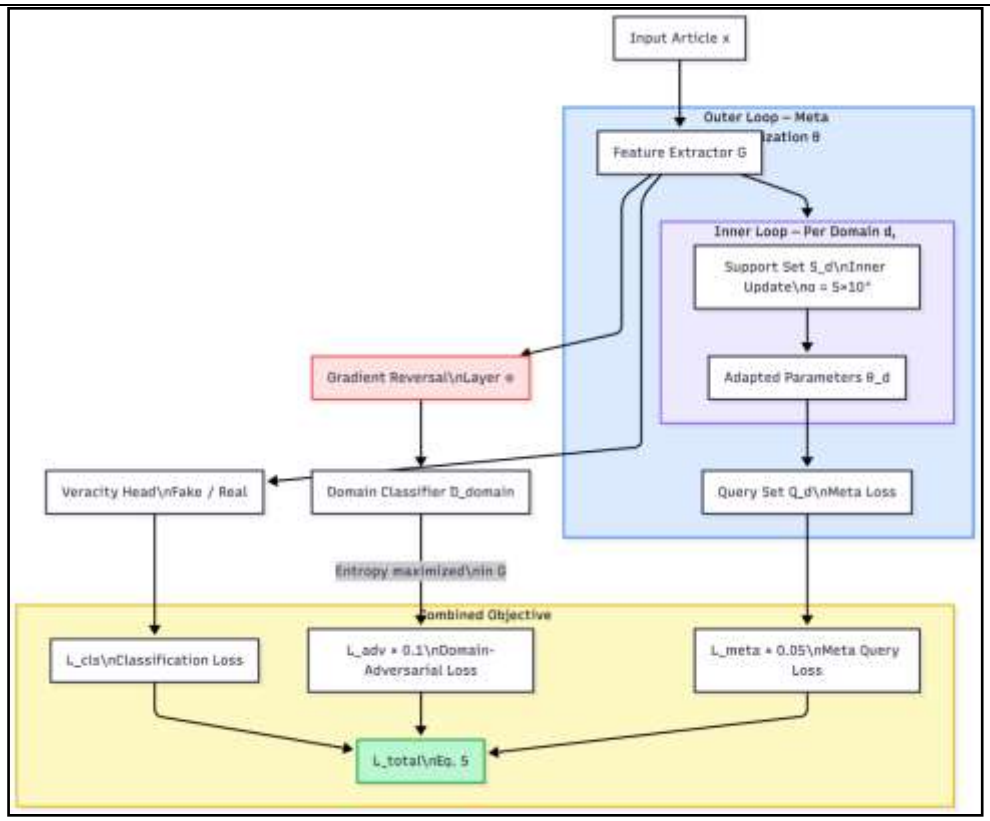

Fig. 2 — DIML Training Schematic

Fig.2 The schematic diagram of DIML training procedure. The shared multi-modal encoder G provides inputs for the veracity module (loss Lcls) and the domain classification (through GRL layer) that trains G for domain-invariance by minimizing Ladv (Equation (3)). Inner gradients for support samples in each domain are updated K times based on Equation (4). Then, Lmeta (Equation (5)) is used for updating outer variables. The meta-optimization MAML algorithm uses loss Lmeta to update the shared parameter vector θ after K iterations using inner loop adaptation of θ to θ′d based on loss Lsupport; at the same time, classification loss Lcls and adversarial loss Ladv are part of the same forward pass. Objective interaction. Every iteration consists of four stages: (i) drawing domains and support/query pairs, (ii) training for K iterations using Lsupport (training classifier on support samples, receiving an adversarial signal through G network), (iii) evaluating Lquery on query samples with respect to θ′d, and (iv) backward pass of Lmeta along with Lcls and Ladv on a per-batch level. While adversarial debiasing modifies representations continuously, the loss Lmeta makes it easier to fine-tune models to individual domains. Three losses mentioned above are combined into Ltotal with hyperparameters $\lambda_{adv} = 0.1$ and $\lambda_{meta} = 0.05$ chosen from the set $\lambda_{adv} \in \{0.01, 0.05, 0.1, 0.5\} \times \lambda_{meta} \in \{0.01, 0.05, 0.1\}$.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

Optional Extension: DIML. Here, we take the complete formulation (6) including the meta term to be optional, applicable to scenarios where quick adaptation using support data is useful. A large value of λmeta compared to λadv may lead to optimization conflicts (intra-loop specialization versus domain confusion), thus we keep λmeta conservative and choose it via validation. Additionally, optional schedule strategies such as +DANN warm-start followed by applying the full meta term may help. If only +DANN (or λmeta = 0) works better (or underfits) than the complete objective, even under matched compute, +DANN (or λmeta = 0) continues to be a strong default benchmark method.

3.2 Uncertainty-guided retrieval, DSRG, and stance

During inference, 30 forward passes of dropout 12 yield mean p̂ and σ2(x) on the veracity head. A fast route (high p̂ and low σ2) avoids deep retrievals. Otherwise, using calibrated σ2 from validation 12,13 (range 0-1) results in retrievals of depth 5, 10, or 20 (Table 2).

Table 2 Retrieval depth vs. calibrated uncertainty (*T 30*).

Calibrated *σ*2	**Documents retrieved**
High (0.8)	20
Medium (0.5–0.8)	10
Low (0.5)	5

DSRG is a three-layer GCN with edge weights $exp(-\Delta t / \tau)$, τ *30* days, set as an engineering prior reflecting typical news-cycle recency; formal τ selection is deferred to future work:

$$H(l+1)\ \sigma(\tilde{D}^{-1/2}\ \tilde{A}\ \tilde{D}^{-1/2}\ H(l)\ W(l)) \quad (7)$$

$\tilde{A} = A + \hat{I}$ is an adjacency matrix including self-loops; $\hat{I}$ is a diagonal degree matrix corresponding to $\tilde{A}$; H(l) includes the node embeddings at layer l (|V| number of nodes and dh hidden size).

Difference between temporal decay and traditional temporal GCNs: While most temporal GCNs used for rumors or diffusion adopt discrete time slices or learnable kernels on a propagation sequence fixed at crawl time, DSRG uses a time-continuous, exponentially decaying function exp(- Δ t / τ) for each evidence-source pair, where τ is constant and not learned in this work. This graph is constructed dynamically using outputs of our retrieval process along with information from the indexed LIAR claims collection and not based on a static social cascade observed offline. This helps avoid weighting stale or reused evidence

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

within the temporal GCN aggregation step, unless the evidence remains highly connected via topology. Stance aggregation takes three components into account:

vscore Σi=1C ci Σj=1ki wij stanceij reliabilityj (8)

With optional ci ≥ 0, Σi ci = 1. The weights are learnt either in an end-to-end way or using two-stage training based on validation evidence loss. Worked example (one inference trajectory). In an example of a query, where the reliable source has a reliability score of 0.732 while an unreliable source has a reliability score of 0.683 for a passage relevance of 0.8, the aggregation performed by DSRG algorithm down-weighs the unreliable source in proportion to its decay/reliability, and the resulting multi-modal veracity module produces p(fake)=0.608 the above values obtained from a single test example.

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

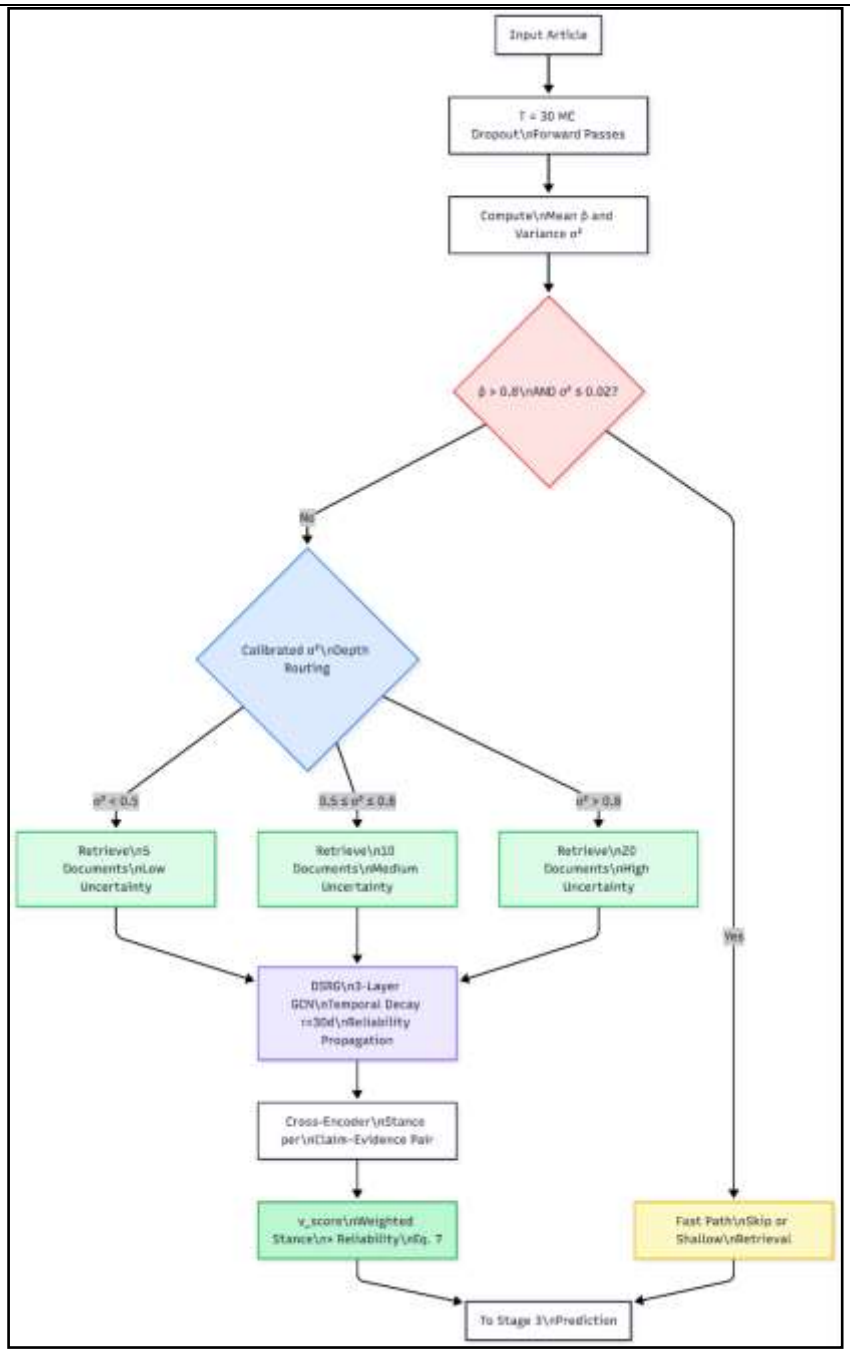

Fig. 3 — Uncertainty-Guided Retrieval Flow

Figure 3 Inference process for uncertainty-driven document retrieval and evidence scoring. The head veracity model is executed T=30 times with dropout activated. In case the average confidence is high and raw variance low, the system takes the fast path route (no deep retrieval). Otherwise, the MC variance is scaled within 0 to 1 range based on the validation data set using min-max scaling and then matched against retrieval depths equaling 5, 10, and 20 retrieved documents (see Table 2). The retrieved passages go through DSRG module

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

(GCN with temporal edge decay, see Eq. (7)). Cross-encoding provides a score for each passage. The score is combined with source reliability into evidence score vscore (see Eq. (8)).

3.3 AI-text ensemble and explanations

The six pre-fixed weights, all non-negative and summing up to 1, yield a unified score between 0 and 1, and the binary decision is obtained by thresholding on 0.55. The weighting system is hardcoded based on our prior belief and thus is not a learned fusion. This particular choice of weights indicates the prior that puts relatively more weight on the detectors that have wide coverage over generators in terms of perplexity or burstiness, while less on those heuristics which are specific to certain tasks (e.g., vocabulary size, repetition rate); we will consider replacing them by validation-optimized or learned weights in future work. Our detectors cover various failure modes of the single-score methods (lexicon perplexity, chunk burstiness, regular expression-based cues, repetition, type/token statistics, and corpus overlap) so that none dominates due to shifts either in domains or generators. Evaluation under paraphrasing attacks is left for future work.

Table 3 Six-detector ensemble (Σk wk 1).

Detector	**Principle**
Perplexity analysis	Pseudo–log-frequency over a closed lexicon
Burstiness analysis	Dispersion across chunks
Linguistic patterns	Regex-style cues
Repetition analysis	N-gram / sentence-start repetition
Vocabulary richness	Type–token-style heuristics
Corpus retrieval (HC3)	Trigram Jaccard vs. reference strings

Expert, moderator, and public explanation levels progressively disclose the same verdict through the API, although we do not report a user study.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

3.4 Experimental harness and data

Benchmarks employ the open source (Python, FastAPI, PyTorch, Transformers; optionally Docker); React front-end interfaces with the API. The veracity experiments run with DistilRoBERTa and classifier head 768 → 256 → 2 with max sequence length 128. Main comparisons are made with equal epoch limits and early-stopping criteria (Table 4b). Training set / validation set / test set sizes are 15,173 / 3,251 / 3,253 (in FakeNewsNet-style), that is, test set size n 3,253.

1. Evidence Retrieval: LIAR-based knowledge base (~12.8K claims): each claim is represented by a sentence embedding with the same DistilRoBERTa encoder and indexed using FAISS; the retrieved claims provide factual groundings for stance evaluation, but not veracity classifications.
2. Evidence Depth: k 5 ⌊u × 15⌋ in the case of disabled Monte Carlo routing, where u 1 − pmax, which means k ∈ 5, 20.
3. Early Exit: applied when maximum class probability pmax ≥ 0.90 in a single forward pass; this forward pass occurs prior to Monte Carlo probe execution; if pmax ≥ 0.90 for the pass, the T 30 Monte Carlo probing is bypassed.

Module-to-script mappings and frozen outputs are described in the repository README.

3.5 Evaluation protocol

The veracity numbers associated with primary runs adhere to the procedure laid out in section 3.4 regarding the fixed test set split (n = 3,253). Mean ± standard deviation across available runs is provided by method and n values per method are noted in the table. Significance testing is not conducted due to differences between paired methods in their respective sets of seeds found within the run archive. Diagnostic metrics for tag-stratification are reported only as difficulty scores (section 4.2). Results for AI-generated-text runs provide accuracy, macro-F1, and AUROC scores at 0.55 threshold on an HC3-style snapshot dataset (Table 9).

4 RESULTS AND ANALYSIS

All the measures were carried out using Google Colab on the NVIDIA Tesla T4 GPU, except when otherwise specified. The evaluation framework is described for (i) domain adaptation

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

(+DANN versus optionally DIML), (ii) uncertainty-based retrieval routing, (iii) DSRG, (iv) AI-text fusion, and (v) stance and explanation paths according to section 3.5.

4.1 Veracity with multi-seed aggregation

Main Result: Table 4b gives the veracity table: mean ± standard deviation of Baseline, +DANN, and +DIML under the protocol in section 3.4; values are averaged over all seeds finished and saved per approach (n in the rightmost column). Differences between methods in their weighted F1 scores are small fractions of a percentage point, smaller than cross-seed standard deviations and less than typical variations between runs for this split. Specifically, there is no statistically significant separation between the mean weighted F1 score of +DANN (84.97%) and the baseline (84.91%), based on overlapping standard deviations (Table 4b). The paper does not consider +DANN to be a standalone improvement in terms of performance; rather, the innovation resides in the open system and reproducibility of the harness itself (section 5). Comparison of DIML to +DANN is made in section 3.1: +DIML yields marginally higher mean weighted F1 score than +DANN, although the former has a greater variance; DIML is a selectable optional module.

Tag diagnostics rely on the baseline checkpoint; see section 4.2. Literature rows (Tables 5–6) provide context for in-domain and OOD literature; they are not direct replications within this codebase.

Table 4b Veracity with multi-seed aggregation (matched epoch cap; primary).

Method	Acc. (%)	Weighted F1 (%)	Macro-F1 (%)	*n* (seeds)
Baseline	85.71 ± 0.23	84.91 ± 0.30	78.57 ± 0.65	4
+DANN	85.70 ± 0.42	84.97 ± 0.19	78.72 ± 0.09	3
+DIML (matched budget)	85.93 ± 0.63	85.31 ± 0.79	79.30 ± 1.29	5

Means ± std over completed Colab runs (Tesla T4); epoch cap matched across methods. Baseline and +DANN use fewer than five seeds in the archived bundle; we do not claim five-seed symmetry or paired tests across all pairs.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

Table 5- Literature F1.

Method	F1	Source
EANN 9	87.1	9
MUSER 4	90.9	4
DAL 3	88.4	3
UMD² 5	85.6	5

Table 6 presents the reported in-domain, out-of-domain, and per-domain weighted F1 scores for DAL, UMD², and MUSER models 3–5. Table 7 splits the bundle file based on whether it was labeled PolitiFact or GossipCop (baseline checkpoint): 59.56% and 75.66% in macro-F1—an intra-baseline diagnostic test, not an inter-baseline test. Previous research has found large gaps between in-domain and OOD performance (by 12 to 16.6 percent points), 3–5 using explicit OOD settings.

Table 6 Published in-domain / OOD / per-domain weighted F1 3–5.

Method	In-domain F1	OOD F1	Δ (pp)	Business	Technology
DAL 3	88.3	76.3	−12.0	75.1	77.4
UMD² 5	85.6	72.8	−12.8	71.2	74.3
MUSER 4	91.1	74.5	−16.6	73.2	75.8

4.2 Train–test domain shift (OOD)

This work does not provide a fully generated multi-seed table where train/test are separated by different domain labels, say, training on GossipCop and testing on PolitiFact, with the same preprocessing as used in Table 4b. Scripts and split builders for these experiments can be found in the open-sourced repository. The tag-stratified macro-F1 scores in Table 7 and Fig. 4 provide an estimation of within-test difficulty by metadata tags for a particular checkpoint as a baseline; it cannot be interpreted as OOD generalization at all. Below is a

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

comparison of macro-F1 by news domain tag against the full-test baseline of that particular checkpoint.

Table 7 Tag-stratified macro-F1 on subsets of bundled test.json (baseline checkpoint).

Subset	Macro-F1 (%)
PolitiFact-tagged	59.56
GossipCop-tagged	75.66

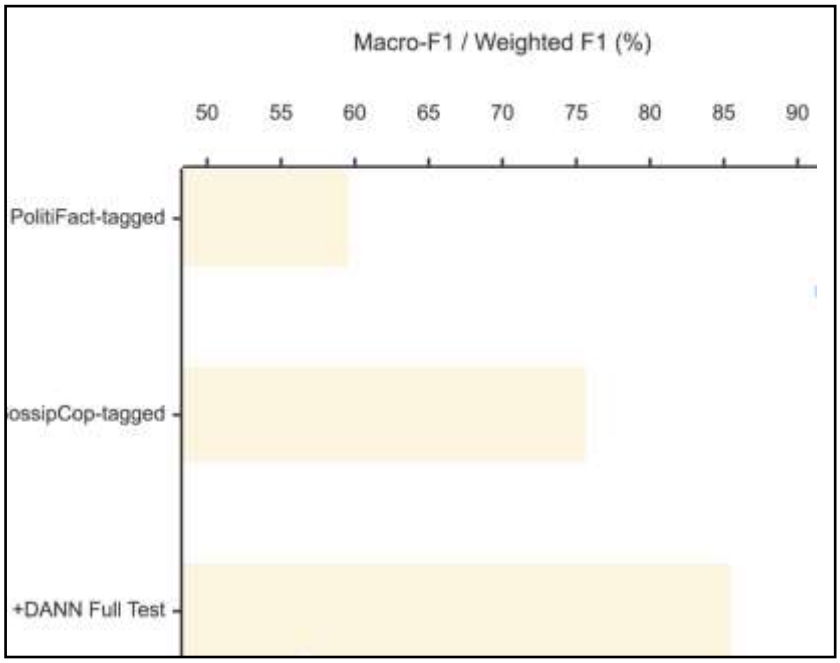

Fig. 4 — Tag-Stratified Diagnostic Bar Chart

Fig. 4 Macro-F1 by news tag vs. full-test baseline (diagnostic). Bars present macro-F1 on PolitiFact-tagged and GossipCop-tagged subsets of the bundled test file for one checkpoint from Table 7. The horizontal dashed line shows the same checkpoint's full-test weighted F1 score (85.22%). This number represents a legacy single-run evaluation on the very split, whereas multi-seed results are provided in Table 4b. Weighted F1 and subset macro-F1 scores are incomparable.

Articles labeled PolitiFact are more challenging than GossipCop articles (59.56% vs. 75.66% macro-F1).

LIAR stance (pipeline diagnostic). The model trained using DistilRoBERTa-head for 3 epochs shows an accuracy of 46.25% and a macro-F1 score of 0.451 on LIAR data (n=1,267,

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

6-way classes 14). Systems using transformers with stronger encoder architectures and greater training time have achieved much higher accuracy on 6-way LIAR classification tasks, but the numbers provided here indicate the accuracy of the default 3 epochs of DistilRoBERTa-head without further optimization. The LIAR numbers are not the state-of-the-art results for stance detection, just what comes with the default API.

4.3 AI-generated text detection (standard benchmarks)

Table 9 gives summary results for the six detector proof-of-concept ensemble on an HC3 style JSONL snapshot 16. Calibration gap. Average AUROC 0.777 (Table 9 footnote) suggests that the fusion score can distinguish between human-written and machine-written lines above chance, yet the accuracy of around 25% at the threshold of 0.55 is indicative of a badly calibrated fusion score distribution at this operating point: there is little mass on either side of the cut-off, leading to collapsed predictions and hence accuracy reflecting the probability of class occurrence. Framing. The AI Text module represents a proxy system-level API that provides hooks to measure time and score fusion, but note that calibration is a minimum pre-condition to using the ensemble in a threshold classification setting: Platt or temperature scaling must be used.

Table 9 AI-text ensemble on HC3-style benchmark.

Benchmark	Acc. (%)	Macro-F1 (%)	AUROC
HC3-style (full)	25.09	20.06	0.777

Note: AUROC is calculated by averaging results from both even-line and odd-line splits of the same source file (0.788 and 0.765), or (0.788 + 0.765) / 2 = 0.777. Accuracy and F1 macro scores indicate problems with thresholding at the 0.55 mark; see Section 4.3.

4.4 Ablations isolating DIML, retrieval, DSRG, and AI-text

The component ablations (MC routing enable/disable, DSRG enable/disable, AI-text fusion, DIML enable/disable) within the same fixed seed grid are not listed above; isolating their effects properly necessitates corresponding configurations of the API and further tests outside the veracity aggregation baseline runs. The repo documents which toggles need service-level experimentation (see section 5, Limitations and future work).

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

4.5 Performance–latency trade-off (Monte Carlo routing)

Measuring on T4 (batch size 1; one Colab run; production deployments may vary—consider as order of magnitude, not a benchmark): encoder forward ≈6ms (mean over 80 runs); end-to-end API route ≈400–500ms or ≈200ms on fast path (pmax ≥ 0.90). The AI stack contributes ≈4 ms per batch of short strings. MC Dropout accounts for most variability of cost: from T 0 to T 30, the forward cost increases from ≈209ms to ≈267ms for long inputs, and from ≈6.4 ms to ≈233ms for short inputs—depth routing thus incurs predictable multiplicative cost to obtain uncertainty-aware retrievals. The increased MC cost on short input is probably due to cold-start overheads in GPU kernel launch on Colab that are smoothed out with longer sequences. For deployments under strict latency constraints, they may consider reducing T (such as T 10) or expanding fast path bands; we encourage them to report (T, latency, F1) triples along with timing logs in the public repo.

5 DISCUSSION AND CONCLUSION

Results in Context: Table 4b represents the core veracity evidence. As indicated in section 4.1, the margins across methods fall below cross-seed noise, and hence cannot be classified as isolated performance boosts. The claim made empirically by this paper is thus not about an isolated performance boost from either +DANN or DIML in this split but is rather about the fact that these two objectives can be trained within one open framework. Tables 5–6 contextualize headline F1 with respect to published work in various experimental settings without claiming any numerical advantage over EANN, MUSER, DAL, and UMD² using this codebase.

Contribution Rethinking: REMIX-FND's core contribution relates to reproducibility at the level of the systems. Specifically, the contribution involves bundling of multi-modal fusion, handling of missing modalities, optionally, training of both DANN and DIML, uncertainty-based retrieval, DSRG, auxiliary AI text and stance scoring, explanations, and a reference API with split files/scripts. This approach is consistent with best practices in documentation for accountable deployment of ML solutions, where the focus is on documented input data, reproducible evaluation harnesses, decoupling of experiments and claims, along with public health guidelines.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

Limitations and future directions: (i) Out-of-domain evaluation: Out-of-domain (OOD) evaluation does not explicitly appear in this paper (section 4.2). Tag subset performance remains diagnostic. (ii) Ablations: A comprehensive component ablation grid using a single seed set is missing. (iii) AI-text: The AI-text route needs to be re-calibrated prior to use (section 4.3); paraphrase attacks 6 have not been benchmarked. (iv) LIAR stance: Accuracy figures refer to the bundled head. LIAR stance state-of-the-art has not been included in the results. (v) Statistics: Statistical comparisons between each method pair have not been provided since the seed sets vary by method in the released bundle. (vi) Stance and veracity coupling: Evaluation in terms of both explanation quality and stance is beyond the scope of this paper.

Future work will include: OOD evaluations following Table 4b style, ablations, more economical uncertainty estimates than T=30 probes, learned or validation-based AI-text fusions, improved LLM-era detectors, and user studies of multi-level explanations.

Ethics: Misinformation detection scores, if utilized without proper oversight, may become tools for censorship and harassment because, due to their biased training data that favors news sources from Anglo-American cultures 1, false positives will have a disproportionate impact on journalists, activists, and people using minority languages. The FakeNewsNet datasets themselves embody cultural and demographic bias in terms of selection and time period; both the graph and retrieval subtasks are prone to popularity bias (i.e., the most-cited sources would seem to be "more reliable"). Geographic and linguistic generalization remains to be seen: the splits provided have an inherent bias towards English, and transferring them to different geographical or language contexts poses open risks. The proper practice in terms of usage includes maintaining human review and appeals process, as well as publishing transparency reports and monitoring for potential biases, which is consistent with WHO guidelines on infodemics 2. Automated decisions should not lead to sanctions on user accounts alone.

REMIX-FND combines multi-modal misinformation scoring, domain adaptation, uncertainty-driven search, temporal source graph representations, and API-ready output in one easy-to-reproduce artifact. The most significant contribution is at the systems level, where matched-budget training and togglable components allow for extension towards OOD testing, calibrated AI vs. text detection, and regional validation without having to rewrite the multi-modal core.

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

ACKNOWLEDGMENT

The authors declare that no financial or institutional support was received for this research.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

REFERENCES:

1. Lazer, D.M.J., Baum, M.A., Benkler, Y., et al.: The science of fake news. Science 359(6380), 1094–1096 (2018). https://doi.org/10.1126/science.aao2998
2. World Health Organization: Managing the COVID-19 infodemic. WHO Technical Report (2020). https://www.who.int/news-room/feature-stories/detail/managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation
3. Zhou, X., Xue, Y., Zafarani, R.: Out-of-distribution evidence-aware fake news detection via dual adversarial debiasing. In: Proc. AAAI, vol. 37, no. 4, pp. 5596–5604 (2023). https://ojs.aaai.org/index.php/AAAI/article/view/25845
4. Liu, Z., Xiong, C., Sun, M., Liu, Z.: MUSER: A multi-step evidence retrieval enhancement framework. In: Proc. ACL, pp. 2791–2803 (2022). https://aclanthology.org/2022.acl-long.194/
5. Chen, Y., Mao, Y., Zhu, H., Ying, Y.: Unsupervised domain-agnostic fake news detection using multi-modal weak signals. In: Proc. ACM MM, pp. 5537–5546 (2022). https://doi.org/10.1145/3503161.3548113
6. Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. Trans. Mach. Learn. Res. (2025). https://openreview.net/forum?id=OOgsAZdFOt
7. Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., Shu, K.: Mining dual emotion for fake news detection. In: Proc. WWW, pp. 3465–3476 (2021). https://doi.org/10.1145/3442381.3450004
8. Qi, P., Cao, J., Yang, T., Guo, J., Li, J.: Exploiting multi-domain visual information for fake news detection. In: Proc. ICDM, pp. 518–527 (2019). https://doi.org/10.1109/ICDM.2019.00062

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/817fqh85

9. Wang, Y., Ma, F., Jin, Z., et al.: EANN: Event adversarial neural networks for multi-modal fake news detection. In: Proc. KDD, pp. 849–857 (2018). https://doi.org/10.1145/3219819.3219903
10. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation. In: Proc. ICML, pp. 1126–1135 (2017). https://proceedings.mlr.press/v70/finn17a.html
11. Ganin, Y., Ustinova, E., Ajakan, H., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17(59), 1–35 (2016). https://www.jmlr.org/papers/v17/15-239.html
12. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proc. ICML, pp. 1050–1059 (2016). https://proceedings.mlr.press/v48/gal16.html
13. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. ICML, pp. 1321–1330 (2017). https://proceedings.mlr.press/v70/guo17a.html
14. Wang, W.Y.: "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proc. ACL, pp. 422–426 (2017). https://aclanthology.org/P17-2067/
15. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: 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 (2020). https://doi.org/10.1089/big.2020.0062
16. Guo, B., Zhang, X., Zhang, Z., et al.: How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv:2301.07597 cs.CL (2023). Accessed: 3 April 2026. https://arxiv.org/abs/2301.07597