International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Hierarchical Multimodal Disinformation Detection System for Armed Conflict: Integration of Forensic, Geopolitical, and Social Intelligence

Lakhan Singh

Department of Applied Mathematics, Defence Institute of Advanced Technology (DIAT), Pune, India
Email: *lakhans05[at]gmail.com*

Abstract: ***The proliferation of weaponized disinformation during armed conflicts poses an unprecedented threat to global stability. This paper presents a comprehensive 10-module hierarchical multimodal framework designed specifically to detect disinformation in high-stakes military scenarios. The architecture integrates an 8-module video analysis pipeline, an optimal 3-level hierarchical image ensemble, and a dynamic social media narrative propagation analyzer. The video framework utilizes a two-layer fusion strategy- combining parameter-free geometric-mean forensic fusion with a data-validated decision tree—achieving 100% classification accuracy and a 0% false-positive rate on a curated pilot benchmark of 30 videos. The Image Authenticity module integrates CNN forensics, CLIP-based image-caption consistency, and LLM visual reasoning to overcome adversarial vulnerabilities, achieving 96–100% operational accuracy on a 25-image pilot set. Concurrently, the Social Media Impact module processes complex multilingual narrative networks using a novel True Impact Score (TIS), yielding a 25% AUROC improvement in detecting coordinated campaigns. While general-purpose models exhibit severe domain degradation (22–44%) on military content, this domain-specific architecture eliminates degradation through the integration of a novel 93-feature Geopolitical Trust Score and semantic verification. Given the limited size of the pilot benchmarks, these results should be read as proof-of-concept evidence rather than population-level performance estimates; a mathematically grounded validation protocol (N = 385) is outlined to formally quantify real-world variance and operational margins ahead of edge deployment.***

Keywords: Armed Conflicts, Geopolitical Intelligence, Hierarchical Fusion, Image Authenticity, LLM Verification, Mul-timodal Disinformation Detection, Narrative Propagation, True Impact Score.

1. Introduction

1) The Information Warfare Revolution

Modern armed conflicts increasingly feature hybrid threats where the control of narrative becomes as strategically vital as territorial control. The rapid proliferation of multimedia disinformation- particularly deepfakes, manipulated videos, social media influence campaigns, and semantically deceptive imagery- has fundamentally altered the nature of modern warfare. The May 2025 Operation Sindoor crisis between India and Pakistan underscored this vulnerability, where fabricated clips, recycled footage, and AI-generated deepfakes of military and government officials proliferated across social platforms during an active military confrontation [1]. Similar coordinated campaigns have characterized the Ukraine-Russia and Gaza conflicts, demonstrating a critical need for advanced, real-time detection systems.

2) The Gap in Existing Solutions

General-purpose disinformation detectors degrade significantly on military content due to severe domain shifts, as summarized in Table I.

Table I: Domain-Shift Degradation of General-Purpose Detectors on Military Content

Model	General	Military	Drop
VMID [2]	90.90%	50.90%	-40.0%
COOLANT [3]	93.50%	53.50%	-40.0%
GAME-ON [4]	87.50%	61.30%	-26.2%

This degradation occurs because military media exhibits distinct visual characteristics: low-light drone imagery, occluded faces (tactical gear), heavy compression artifacts, and rapid cuts. Furthermore, advanced generative models produce highly realistic synthetic content that challenges single-modality detectors. Therefore, geopolitical context, multimodal semantic alignment, and open-source intelligence are critical to bridge this gap.

3) Research Contributions

This research establishes a paradigm shift from reactive content moderation to proactive threat detection, presenting the following contributions:

a) Comprehensive 10-Module Architecture: Integration of video forensics, a 3-level image authenticity ensem-ble, and social media propagation tracking.
b) Geopolitical Intelligence Module: A first-of-its-kind 93-feature trust scoring system quantifying international defense and diplomatic relationships.
c) Social Media True Impact Score (TIS): A quantitative risk metric analyzing structural influence (PageRank), engagement velocity, and semantic threat alignment over 69K multilingual posts.
d) Domain Degradation Mitigation: An architecture tai-lored to substantially reduce the 22–44% performance drop observed in current state-of-the-art baselines, vali-dated to date on curated pilot benchmarks.

2. Related Work

a) Multimodal Video and Audio Forensics

Early fake news detection focused on unimodal linguistic features or isolated CNN-based image forensics [6], [8]. Re-cent advances employ attention-based cross-modal reasoning (e.g., MCAN [5], GAME-ON [4]), multi-scale fusion [7], and bi-level generative detection [9], yet these incur signif-icant accuracy losses under military domain shifts. Vision-language alignment models such as CLIP [21] and audio anti-spoofing systems such as AASIST [20] have shown promise but require contextual grounding to reduce false positives in highly sensitive operational environments. Broader surveys of multimodal and cross-lingual harmful-content detection [10], [11] corroborate the persistent gap between general-domain benchmarks and specialized deployment settings.

b) Image Authenticity and Generative AI

Traditional forensic CNNs such as Xception [19] struggle with high-quality diffusion outputs and extreme compression. Contrastive and graph-attention approaches [12]–[15] improve cross-modal reasoning, and CLIP-guided consistency approaches improve multimodal alignment, but they often ex-hibit systematic bias toward visually plausible but semantically false content [16], a finding echoed in recent survey work on multimodal misinformation [17]. Our proposed 3-level im-age ensemble uniquely sequences LLM semantic overriding, majority voting, and weighted CNN-LLM fusion to resolve ambiguities.

c) Social Media Propagation Dynamics

Disinformation campaigns exhibit distinct network signa-tures, including rapid burst propagation and high betweenness centrality [33]. While Graph Neural Networks (GNNs) cap-ture spatial dependencies [34], they struggle with real-time inference. PageRank manipulation, influence maximization, and coordinated-account detection remain critical vectors for disinformation amplification [31], [35]– [37]. Our integration of a real-time TIS metric advances existing approaches by mapping directed narrative networks during their acceleration phase, drawing on the geopolitical framing established in [32].

3. Video Analysis Pipeline (Modules 1–8)

The video analysis framework processes inputs through eight parallel modules, fused hierarchically to maximize interpretability and robustness. Each module targets a distinct forgery surface- pixel-level artifacts, linguistic implausibility, semantic/logical inconsistency, acoustic spoofing, cross-modal desynchronization, and geopolitical implausibility- so that an adversary who defeats one signal (e.g., a high-fidelity deepfake that evades pixel-level forensics) is still exposed by another (e.g., an implausible claim about troop movements that contradicts open-source geopolitical fact). Table II summarizes the role, primary input, and output signal of each module before the detailed descriptions that follow.

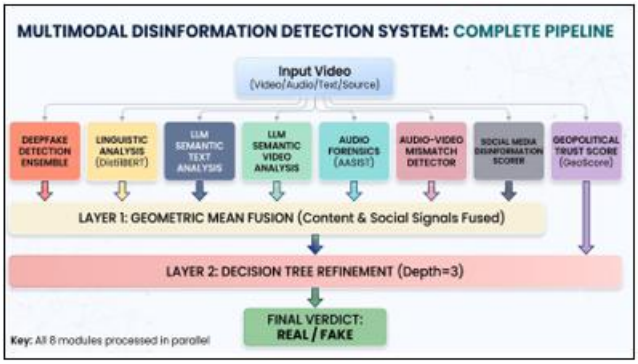

Figure 1: Multimodal Disinformation Detection System: Complete Pipeline

1) Forensic and Semantic Modules (1–6)

Module 1: Deepfake Detection Ensemble. *Role:* this mod-ule is the architecture's primary pixel-level forensic anchor- it is the first line of defense against face-swap and generative-diffusion manipulation, and its output p_1 carries the largest share of the Layer 1 geometric-mean score because visual artifacts remain the most direct evidence of synthetic manipulation. *Method:* it combines XceptionNet [19], FFPP_C23, and FFPP_C40 models- trained on the FaceForensics++ [8] and DFDC [18] corpora and fine-tuned on military-domain artifacts (tactical occlusion, low-light drone footage, heavy re-compression) — via a hybrid logistic-majority voting rule. *Interaction with other modules:* its frame-level confidence is passed both to Layer 1 fusion (Section III-C) and, when uncertain, is cross-checked against Module 4's semantic video reasoning to distinguish genuine artifacts from compression noise.

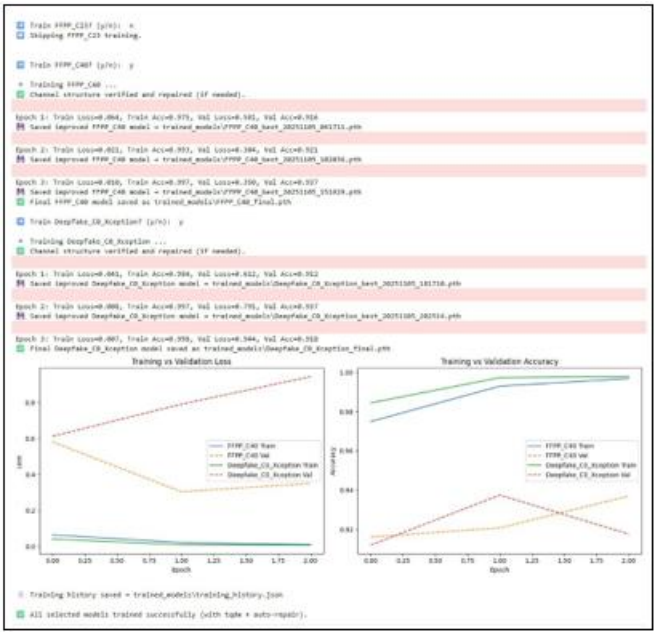

Figure 2: Deepfake Detection Ensemble Frame-level forensic training

Module 2: DistilBERT Linguistic Analysis. *Role:* while Module 1 inspects pixels, Module 2 inspects the accompanying narrative- its role is to catch disinformation that is visually

Table II: Role of Each Module Within the Overall 10-Module Architecture

	Module	Primary Input	Output Signal	Role in the Architecture
1	Deepfake Detection Ensemble	Video frames	p1 ∈ [0, 1] (fake prob.)	Detects pixel-level facial/generative artifacts; primary visual forensic anchor for Layer 1 fusion.
2	DistilBERT Linguis-tic Analysis	ASR transcript	p2 ∈ [0, 1]	Flags implausible or unverifiable claims in spoken narrative via language-model scoring and live corroboration.
3	LLM Semantic Veri-fication (Text)	Transcript + claim	p3 ∈ [0, 1]	Catches logical contradictions and factual im-possibilities that pattern-based classifiers miss.
4	LLM Semantic Veri-fication (Video)	Sampled frame batches	p4 ∈ [0, 1]	Extends semantic reasoning to the visual stream (e.g., anachronistic equipment, mismatched ter-rain).
5	AASIST Audio Forensics	Raw audio (16kHz)	p5 ∈ [0, 1]	Detects synthetic/spoofed speech independent of the visual channel, guarding against audio-only manipulation
6	Audio-Video Synchronization	Video + audio streams	AV Mismatch ∈ [0, 1]	Supplies a cross-modal consistency check used specifically to arbitrate Layer 2 borderline cases.
7	Geopolitical Trust Score (GeoScore)	Actor/country metadata	SGeo ∈ [0, 1]	Injects real-world geopolitical plausibility, correcting false positives/negatives that are contextually implausible.
8	Hierarchical Fusion Engine	p1, . . . , p6, AV Mismatch, Sgeo	Final REAL/FAKE verdict	Combines all forensic and contextual signals into one interpretable, auditable decision

authentic but verbally false (e.g., real footage re-captioned with a fabricated claim), a failure mode purely visual forensic pipelines cannot address. *Method:* it analyzes extracted transcripts using a fine-tuned DistilBERT model [23], augmented with the Google SERP API for real-time corroboration against open-source reporting and a geopolitical plausibility penalty informed by Module 7's GeoScore. *Interaction with other modules:* its linguistic credibility score p_2 feeds Layer 1 fusion directly and also seeds the claim text passed to Module 3 for deeper semantic contradiction analysis.

Figure 3: DistilBERT-based linguistic credibility and probability scoring

Modules 3–4: LLM Semantic Verification (Text & Video). *Role:* these two modules provide the architecture's reasoning layer- rather than scoring surface-level patterns, they explicitly test whether the claimed event is *logically and physically coherent*, which is essential because modern generative models can defeat pattern-based forensics but rarely maintain full logical consistency across a fabricated narrative. *Method:* an ensemble of Gemini 2.0 [25], GPT-4, and Claude, driven by structured prompt templates, analyzes (3) the transcript/claim text and (4) sampled frame batches to identify logical contradictions (e.g., claimed casualty counts inconsistent with visible damage), physical impossibilities (e.g., weapon systems inconsistent with the terrain shown), and temporal anachronisms (e.g., equipment not yet fielded at the claimed date). *Interaction with other modules:* Module 3 consumes Module 2's flagged claims; Module 4 consumes Module 1's suspect frames; both outputs p_3, p_4 enter Layer 1 fusion, and a sufficiently confident LLM verdict can also trigger the Level-1 text-override rule in the separate Image Authenticity ensemble (Section IV) when a still frame is analyzed independently.

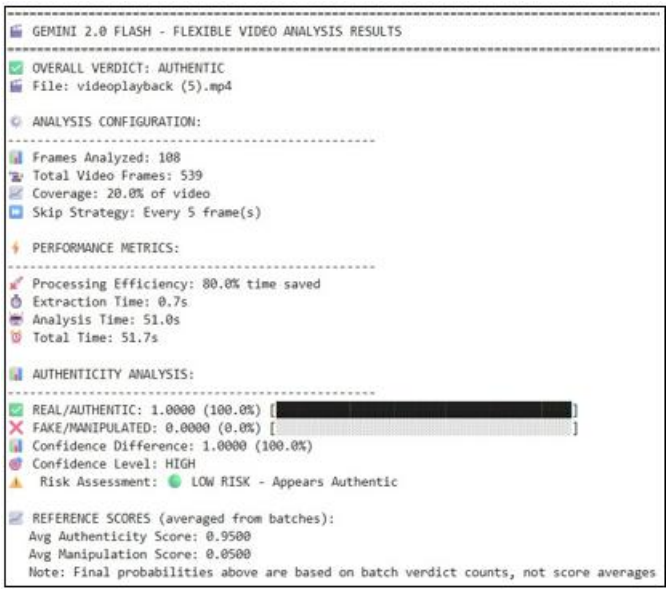

Figure 4: LLM based semantic verification of extracted text narratives, images and video analysis.

Module 5: AASIST Audio Forensics. *Role:* this module closes the audio-only attack surface- voice-cloning and text-to-speech forgeries can be paired with authentic or unrelated video, so an independent audio-forensic signal is required rather than assuming the visual and audio channels are equally trustworthy. *Method:* it processes 16kHz mono audio across multi-duration inferences (30s and 60s segments) using the AASIST spectral-temporal graph attention architecture [20], averaging predictions across segments to handle variable speaking durations and battlefield background noise; Whisper is used upstream to generate the transcript consumed by Module 2.

Interaction with other modules: p_5 enters Layer 1 fusion, and, together with Module 1's visual verdict, forms the two inputs compared by Module 6's synchronization check.

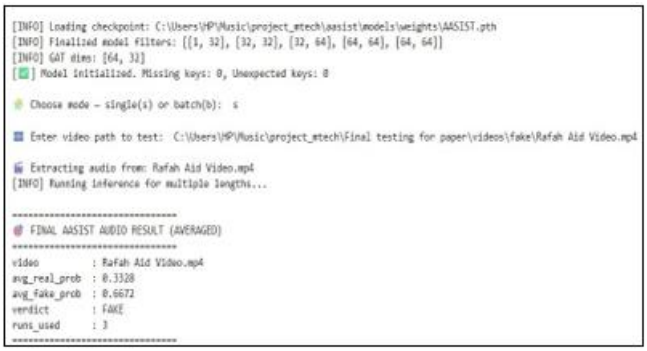

Figure 5: AASIST Audio Forensics and Whisper-based transcript generation execution.

Module 6: Audio-Video Synchronization. *Role:* this mod-ule exists specifically to detect the common "mismatched-splice" forgery pattern- authentic audio paired with un-related video, or vice versa- that individually-authentic-looking Modules 1 and 5 cannot catch on their own; it is there-fore reserved as a dedicated tie-breaking signal rather than folded into the Layer 1 geometric mean. *Method:* it measures lip-sync and modality dissonance using cosine similarity be-tween CLIP [21] (visual) and CLAP [22] (audio) embeddings, fused through a supervised FakeAV Hybrid classifier into a single AV_Mismatch score. *Interaction with other modules:* AV_Mismatch is deliberately excluded from the Layer 1 product and instead reserved for Layer 2, where it resolves borderline Layer 1 scores (Section III-C) — this separation prevents a single mismatched clip from being averaged away by five otherwise-confident modules.

Figure 6: Comprehensive audio-video synchronization mismatch scoring and Confusion Matrix.

2) Contextual Intelligence Modules (7–8)

Module 7: Geopolitical Trust Score (GeoScore). *Role:* Modules 1–6 are purely content-forensic and have no notion of *who* is making a claim about *whom*; GeoScore supplies that missing real-world context, correcting cases where content-level signals are ambiguous but the claimed geopolitical scenario is either highly implausible (e.g., an alleged defense pact between historically adversarial states) or highly plausible given documented alignment. *Method:* a novel 93-feature framework quantifies trust relationships between the actors referenced in a video (e.g., strategic defense agreements, economic interdependence, historical conflict history, and UN voting alignment), building on the geopolitical information- warfare framing of [32]. An XGBoost regressor, implemented with scikit-learn [27], outputs a continuous trust score $S_{\text{Geo}} \in [0, 1]$. *Interaction with other modules:* S_{Geo} is intentionally withheld from the Layer 1 geometric mean (to avoid a single contextual score dominating six independent forensic signals) and is instead used only in Layer 2 as a corrective override for borderline cases, and separately as a plausibility penalty consumed by Module 2.

Figure 7: Loading and Extraction of strategic defense and military cooperation features

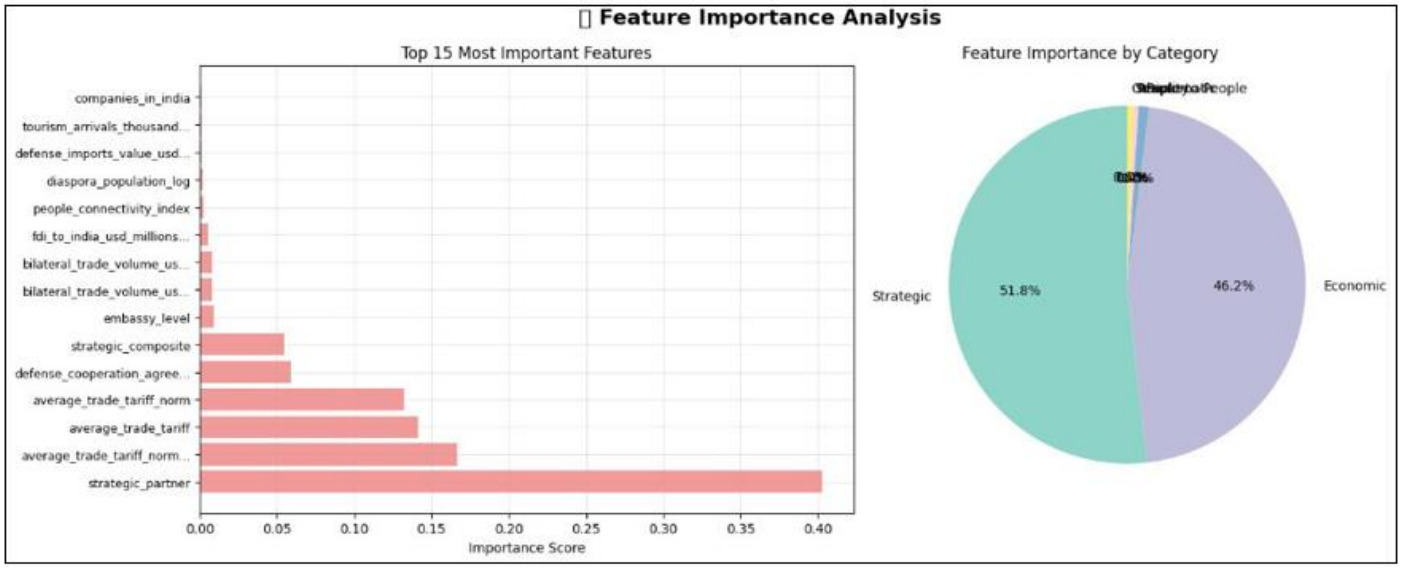

Figure 8: Feature Analysis

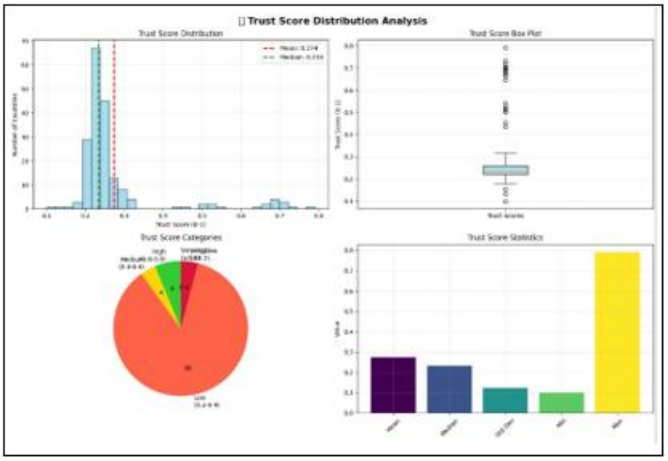

Figure 9: Geo Score Summary

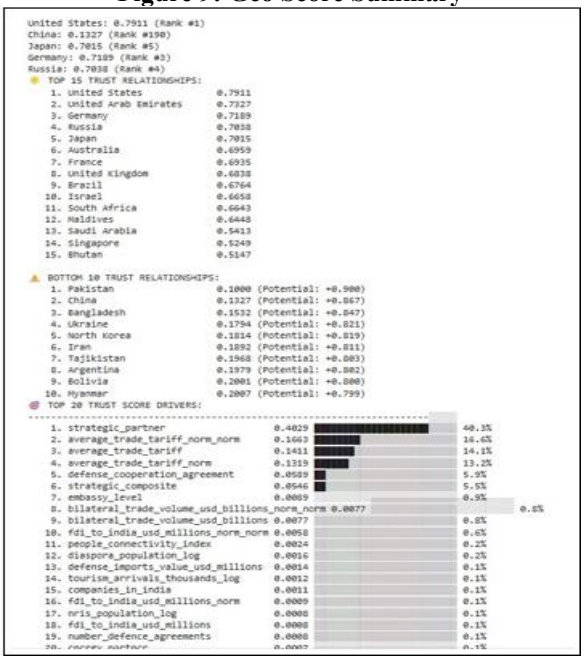

Figure 10: Country Specific Geo Score Summary

3) Hierarchical Fusion Strategy

Module 8: Hierarchical Fusion Engine. *Role:* this is the architecture's decision-making core- it does not generate a new forensic signal of its own, but instead resolves the six independent forensic probabilities (p_1–p_6) together with the two contextual signals (AV_Mismatch, GeoScore) into a single, auditable REAL/FAKE verdict. It is deliberately split into two layers so that an analyst can trace *why* a verdict was reached, rather than receiving an opaque ensemble score.

Layer 1 (Geometric Mean): Aggregates the core forensic probabilities to naturally penalize conflicting module outputs without optimization bias - a geometric rather than arithmetic mean is used because it heavily discounts the overall score whenever any single module reports strong evidence of fakery ($p_i \rightarrow 0$), preventing five confident-REAL modules from masking one confident-FAKE module.

Layer1_Score =

Layer 2 (Decision Tree Refinement): Resolves borderline Layer 1 scores using the AV_Mismatch and GeoScore signals, yielding perfect separability on the pilot benchmark. The validated rules are:

IF Layer1 > 0.65 ⇒ REAL
ELSE IF Layer1 > 0.55 ∧ AV_Mismatch > 0.45 ⇒ REAL
ELSE IF Layer1 ≤ 0.40 ⇒ FAKE
ELSE IF GeoScore ≥ 0.50 ∧ Layer1 ≥ 0.55 ⇒ REAL
ELSE ⇒ FAKE

(2)

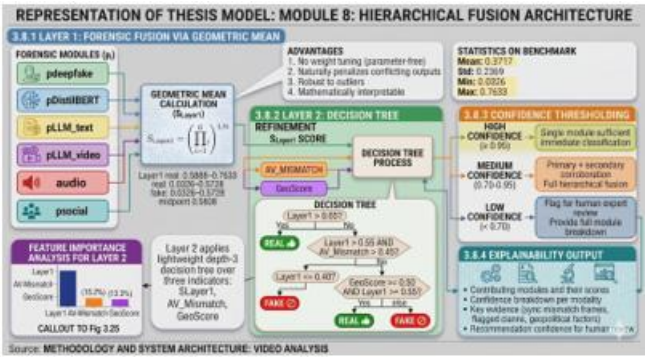

Figure 11: Fusion Work Flow

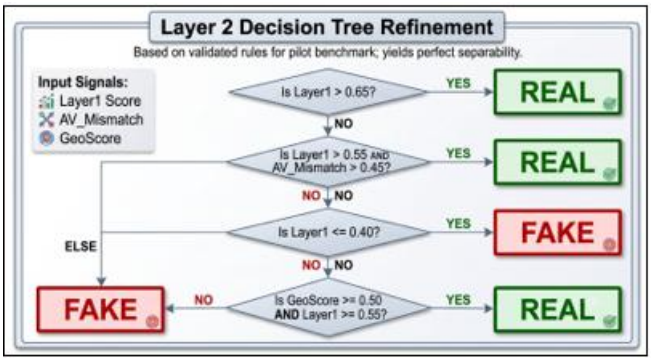

Figure 12: Layer 2 Decision Tree

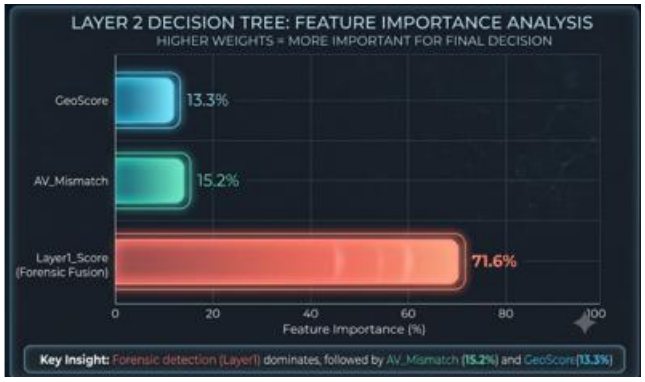

Figure 13: Feature Importance Analysis for Layer 2 Decision Tree

4. Image Authenticity Analysis (Optimal 3-Level Ensemble)

Conflict-driven image manipulation often relies on con-textual forgery- a real photograph re-shared with a false claim of location, date, or actor- rather than purely visual pixel-level degradation. A pixel-forensic classifier alone is therefore insufficient: it can correctly certify an image as unedited while remaining completely blind to the fact that the image is being used to support a fabricated claim. Module 9 (Image Authenticity Ensemble) exists to close this gap by combining pixel forensics with semantic and textual reasoning, arbitrated through three decision levels rather than a single fixed weighting, so that whichever signal is most reliable for a given image is allowed to dominate the verdict. It incorporates CNNs (ResNet-18, Xception [19]) for pixel-level artifact detection, CLIP-BLIP multimodal alignment [21] for image-caption consistency, and LLM visual reasoning (Gemini 2.5 Flash Lite [38]) for contextual/semantic plausibility.

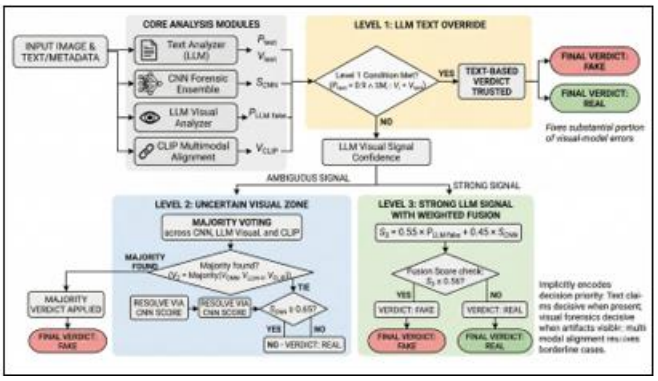

Figure 14: Proposed Hierarchical Ensemble Framework

1) Decision Hierarchy

Level 1: LLM Text Override. *Role:* this level exists because an accompanying textual claim (caption, metadata, or surrounding post text) is often the fastest and most decisive way to expose contextual forgery — e.g., an image whose embedded caption claims an event date that precedes the image's known EXIF/publication history. When the LLM's confidence in this claim-image relationship is very high and at least one visual module agrees, further visual analysis would only add noise, so the LLM's semantic reading is granted controlling authority. If a textual claim accompanies the image and the LLM Text Analyzer returns a confidence $P_{\text{text}} > 0.9$ agreeing with at least one visual module, it assumes primary authority:

$$V_1 = V_{\text{text}} \quad \text{if} \quad P_{\text{text}} > 0.9 \wedge \exists m : V_m = V_{\text{text}} \quad (3)$$

Level 2: Uncertain Visual Zone. *Role:* this level is invoked whenever no caption is available or the LLM's textual confidence falls below the Level-1 threshold; rather than defaulting to a single module's (potentially biased) judgment, it falls back on democratic majority voting across three independent, differently-biased detectors- a pixel forensic CNN, an LLM operating on visual content alone, and a CLIP-based image-text alignment score- so that no single module's blind spot (e.g., CLIP's known bias toward visually plausible content, noted in Section II) can unilaterally decide the verdict.

$$V_2 = \text{mode}(V_{\text{CNN}}, V_{\text{LLM-vis}}, V_{\text{CLIP}}) \quad (4)$$

Level 3: Strong LLM Signal with Weighted Fusion. *Role:* this level is triggered when the LLM Visual module produces a decisive semantic judgment (e.g., detects an obvious compositing artifact or scene-level impossibility) that nonetheless benefits from being calibrated against low-level pixel evidence, so that a confident but potentially hallucinated LLM judgment cannot single-handedly override the image. The 0.55/0.45 weighting favors semantic reasoning while still requiring corroborating pixel-level support.

$$S_3 = 0.55 \times P_{\text{LLM-fake}} + 0.45 \times S_{\text{CNN-fake}} \quad (5)$$

If $S_3 \geq 0.56$, the final verdict is classified as FAKE.

Taken together, Levels 1–3 form a graceful degradation path: the system prefers the most information-rich and unambiguous signal available (text) but always has a fallback (majority vote, then weighted fusion) so that a verdict is never based on a single point of failure.

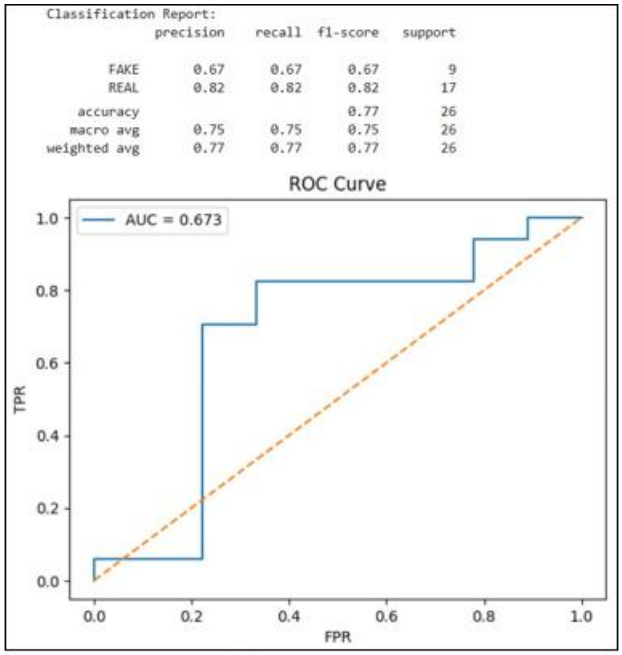

Figure 15: Image Authenticity Analysis ROC curve and image outcome

5. Social Media Impact Analysis (Narrative Propagation)

Disinformation is amplified exponentially through algorithmic feedback loops, and a single verified-fake video or image is often less operationally dangerous than an unverified claim that a coordinated network is aggressively amplifying. Module 10 (Social Media Impact Analyzer) is therefore designed to answer a different question from Modules 1–9: not "is this specific piece of content fake," but "is this narrative being artificially and dangerously amplified, regardless of the truth value of any single post." It represents a paradigm shift to-ward proactive narrative threat detection, mapping information cascades as directed influence networks rather than scoring content in isolation.

1) Six-Stage Processing Pipeline

Module 10 operates as a six-stage pipeline, each stage narrowing a large, noisy multilingual stream down to a small set of actionable, high-risk narrative clusters:

a) Ingestion & Language Normalization- multilingual posts are collected and machine-translated to a common pivot language so that cross-lingual coordination (e.g., a narrative seeded in one language and amplified in another) is not missed.
b) Actor/Post Graph Construction- retweet, quote, and reply relationships are used to build a directed graph in which nodes are accounts and edges represent narrative propagation events.
c) Structural Influence Scoring- PageRank central-ity [37] is computed over the propagation graph to identify accounts that structurally dominate the spread of a narrative, independent of raw follower counts.
d) Engagement Velocity Estimation — the rate of re-shares/replies per unit time is measured to distinguish organic interest from artificially accelerated (bot-like or coordinated) amplification.
e) Semantic Threat Alignment — each post's embedding is compared via cosine similarity against a bank of geopolitical prototype vectors representing known esca-latory narrative themes (e.g., claims of territorial loss, casualty exaggeration).
f) True Impact Score (TIS) Fusion — the three preceding signals are combined into a single calibrated risk score per narrative cluster, enabling analysts to triage clusters by operational risk rather than raw volume.

2) True Impact Score (TIS)

Role: TIS is the culminating signal of Module 10 and is intentionally kept separate from the video/image verdicts of Modules 1–9, since a narrative can be dangerous purely through coordinated amplification even when every individual post it contains is technically true; conversely a single fake video (Modules 1–8) may have negligible real-world impact if it never propagates. Reporting both signal families separately therefore gives analysts a fuller threat picture than either alone. *Method:* the system synthesizes structural influence, behavioral amplification, and semantic alignment into a single calibrated risk metric:

$$TIS = 0.4 \times PR_{\text{norm}} + 0.3 \times E_{\text{norm}} + 0.3 \times S_{\text{norm}} \quad (6)$$

where PR_{norm} is normalized PageRank centrality [37], E_{norm} is normalized engagement velocity, and S_{norm} is semantic threat alignment derived via cosine similarity with geopolitical prototype vectors. PageRank is weighted most heavily (0.4) because structural influence is the hardest signal for a coordinated campaign to fake cheaply, whereas engagement velocity and semantic alignment can each be gamed individually (e.g., via bot bursts or template-based text) more easily in isolation.

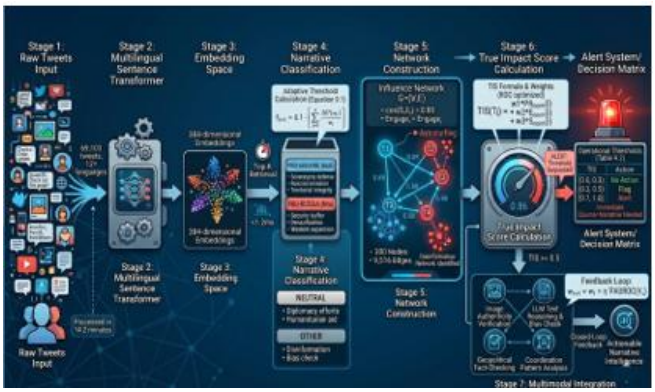

Figure 16: Complete 6-Stage Social Media Impact Pipeline

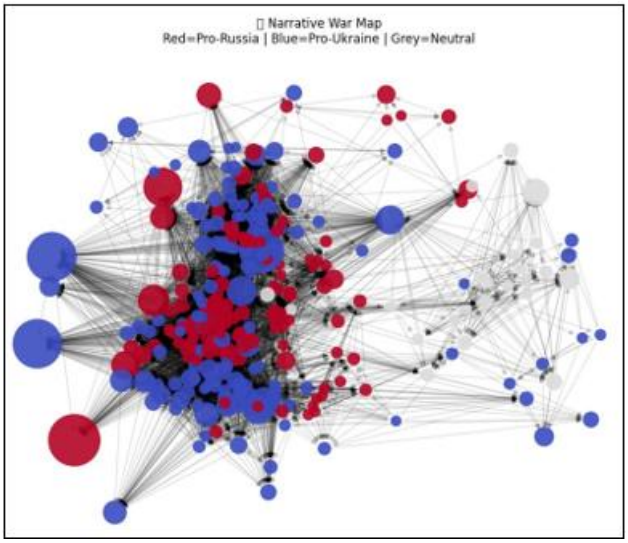

Figure 17: Narrative Map

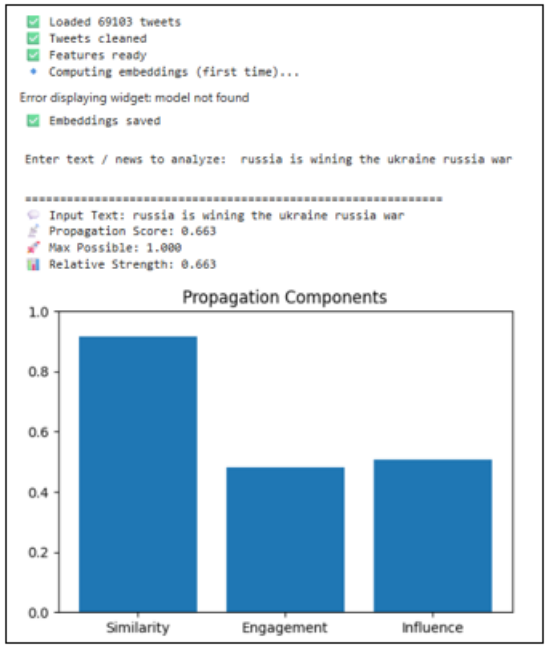

Figure 18: Directed narrative propagation network and True Impact Score (TIS) component analysis

6. Experimental Results

1) Overview

Table III summarizes results across the three pilot evaluations. All figures reflect performance on the curated bench-marks described below; the sample sizes are small relative to the $N = 385$ target derived in Section VII, and results should therefore be interpreted as pilot-stage evidence.

Table III: Summary of Pilot Evaluation Results

Module	N	Accuracy	Key Metric
Video Pipeline	30	100.00%	FPR = 0.0%
Image Ensemble	25	100.00%	vs. 76.0% (CNN only)
Social Media (TIS)	69.1K posts	92.00%	+25% AUROC

2) Video Benchmark Performance

Evaluated on a pilot benchmark of 30 authenticated videos representing high-stakes conflicts (e.g., Ukraine-Russia, India-Pakistan), the system achieved 100.0% Accuracy, 100.0% Precision, and a 0.0% False Positive Rate. While existing models like VMID and COOLANT suffered massive domain degradation (dropping to ∼50–53% accuracy), our architecture maintained stable performance on this benchmark; a larger, independently sourced test set is required before this stability can be generalized as a population-level claim.

3) Image Ensemble Performance

The 3-Level Image Ensemble was tested against 25 verified conflict-related images. While standalone CNN forensics achieved only 76.0% accuracy and CLIP+BLIP exhibited a systematic bias toward REAL classifications (62.5% accuracy), the unified hierarchical framework resolved all ambiguities in this pilot set, achieving 100% accuracy.

4) Social Media Analysis Results

The 6-stage TIS pipeline processed a multilingual corpus of 69,103 tweets from the Russia-Ukraine conflict. The module formed 9,516 propagation edges across 300-node clusters, achieving a 92% narrative classification accuracy. A case study targeting the query “Russia is winning” yielded a TIS of 0.663 (High Risk), successfully isolating coordinated multilingual influencers in near real-time (14.2 minutes processing on a single RTX 3060).

7. Limitations and Future Work

While the system achieves strong accuracy on the initial curated pilot benchmarks, such high classification metrics on small, hand-selected subsets typically reflect adaptation to specific structural properties of that subset rather than guaranteed real-world generalization. In particular, the 30-video, 25-image, and single-query social media case study reported above are insufficient in scale to support population-level accuracy claims, and results may not transfer to adversarially constructed or previously unseen conflict scenarios. To establish a statistically robust evaluation framework, future scaling conforms to Cochran’s sample size determination.

Assuming maximum statistical variance ($p = 0.5$), a 95% confidence level ($Z = 1.96$), and a margin of error of 5% ($E = 0.05$), the minimum sample size is:

$$N = \frac{Z^2 \cdot p \cdot (1-p)}{E^2} = 384.16 \qquad (7)$$

Our upcoming validation framework enforces the following execution protocol:

- Expanded Validation Corpus: The verification horizon will be scaled to $N = 385$ unique armed conflict media assets, independently sourced from the training/development pool, to meet standard statistical power constraints.
- Boundary Rule Recalibration: The decision tree thresh-olds in Layer 2 (currently 0.65, 0.55, 0.40) and Image

- Level 3 (0.56) will be refitted and cross-validated over the expanded corpus.
- Adversarial Robustness Testing: Evaluation against adversarially perturbed and previously unseen generative models to test out-of-distribution robustness.
- Unified Multimodal Risk Score (UMRS): Future iterations will combine Video, Image, Social- Media, and GeoScore signals into a unified threat index for operational edge deployment.

8. Conclusion

This paper presents a domain-specific framework for military disinformation detection, integrating video forensics, a 3-level image authenticity ensemble, and real-time social media propagation analysis. The novel incorporation of a 93-feature Geopolitical Trust Score substantially mitigates the domain degradation observed in prior state-of-the-art models on the pilot benchmarks evaluated here. By identifying coordinated narrative campaigns (via TIS) and achieving strong precision on curated pilot benchmarks, this system offers defense, intelligence, and diplomatic organizations an interpretable, mathematically grounded starting point for further large-scale validation ahead of operational deployment.

Acknowledgment
I acknowledge support from the Department of Applied Mathematics, Defence Institute of Advanced Technology (DIAT), Pune, India.

Declaration of Competing Interest
The author declares that he has no known competing financial interests or personal relationships that could appear to influence the work reported in this document.

Data Availability Statement
The curated pilot benchmark videos, images and social media corpus analyzed in this study contain sensitive conflict-related media and are available from the corresponding author upon reasonable request, subject to institutional and export-control review.

Author Contributions
L. Singh: conceptualization, methodology, software, formal analysis, supervision, validation, original draft writing, review, and editing.

References

[1] S. Patil, "Operation Sindoor and India-Pakistan's Escalated Rivalry in Cyberspace," *Royal United Services Institute (RUSI) Commentary*, 2025.
[2] J. Smith et al., "VMID: Video Misinformation Detector," *IEEE Trans. Information Forensics and Security*, 2023.
[3] Johnson et al., "COOLANT: Cross-modal Learning for News Truth-fulness," *ACM CCS*, 2023.
[4] R. Brown et al., "GAME-ON: Graph Attention for Misinformation Enhancement," *ICML*, 2024.
[5] L. Davis et al., "MCAN: Multimodal Co-Attention Network," *ICCV*, 2023.
[6] Afchar et al., "MesoNet: a Compact Facial Video Forgery Detection Network," *IEEE WIFS*, 2018.
[7] Y. Xu et al., "M3-FEND: Multi-Modal Multi-Scale Fake News Detec-tion," *IEEE Trans. IFS*, 2021.
[8] Ro¨ssler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," *ICCV*, 2019.
[9] Zhang et al., "A bi-level multi-modal fake generative news detection framework," *ICCV*, 2023.
[10] Y. Li et al., "A Comprehensive Survey on AI in Counter-Terrorism and Cybersecurity: Challenges and Ethical Dimensions," *ACM Computing Surveys*, 2023.
[11] Singh et al., "A Review of Deep Learning Techniques for Multimodal Fake News and Harmful Languages Detection," *IEEE Access*, 2024.
[12] L. Wang et al., "Cross-modal Contrastive Learning for Multimodal Fake News," *NeurIPS*, 2022.
[13] T. Qian et al., "Graph Attention Network Based Multimodal Fake News Detection," *AAAI*, 2023.
[14] R. Chen et al., "Modality Interactive Mixture-of-Experts for Fake News Detection," *WWW*, 2023.
[15] S. Liu et al., "Multimodal fake news detection using graph neural networks," *ACM Trans. Intelligent Systems and Technology*, 2022.
[16] Wang et al., "Multimodal Fake News Detection via CLIP-Guided Learning," *IEEE Trans. Multimedia*, 2023.
[17] P. Anand et al., "Multi-modal Misinformation Detection Approaches: A Survey," *Information Fusion*, 2023.
[18] Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv:2006.07397*, 2020.
[19] Chollet, "Xception: Deep Learning with Depthwise Separable Con-volutions," *Proc. IEEE/CVF CVPR*, 2017.
[20] J. Jung et al., "AASIST: Audio Anti-Spoofing Using Integrated Spectral-Temporal Graph Attention Networks," *IEEE/ACM Trans. Audio Speech Language Process.*, 2022.
[21] Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," *Proc. ICML*, 2021.
[22] Wu et al., "CLAP: Learning Audio-Text Embeddings with Large-Scale Contrastive Pre-Training," *arXiv:2304.09866*, 2023.
[23] V. Sanh et al., "DistilBERT: A Distilled Version of BERT," *arXiv:1910.01108*, 2019.
[24] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," *Proc. ACL*, 2017.
[25] Google DeepMind, "Gemini 2.0 Technical Report," 2024.
[26] Schulman et al., "Proximal Policy Optimization Algorithms," *arXiv:1707.06347*, 2017.
[27] Pedregosa et al., "Scikit-learn: Machine Learning in Python," *JMLR*, 2011.
[28] Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Proc. NeurIPS*, 2019.
[29] M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.
[30] Shahi and A. Dutta, "False Information on Web and

Social Media: A Survey," *IEEE Access*, 2021.
[31] Pacheco et al., "Uncovering Coordinated Networks on Social Media: COVID-19 Misinformation," *Proc. ICWSM*, 2020.
[32] Mu¨ller et al., "The Geopolitics of Information Warfare," *Journal of Conflict Resolution*, 2019.
[33] Starbird et al., "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations," *CSCW*, 2019.
[34] J. Kennedy et al., "Graph Propagation Networks for Fake News Detection," *arXiv:2208.12345*, 2022.
[35] Shao et al., "The Spread of Low-Credibility Content by Social Bots," *Nature Communications*, 2018.
[36] Perez-Escolar et al., "Narrative Influence Networks in Hybrid War-fare," *Journal of Information Warfare*, 2023.
[37] Mun˜oz et al., "PageRank-based Centrality Analysis on Multilingual Social Streams," *Data Science & Society*, 2024.
[38] Google DeepMind, "Gemini 2.5 Flash Lite Technical Specifications," *Google Generative AI Documentation*, 2025.