IJRDET

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 01, January 2026)

SHIELD: A Multi-Layered AI Defense Framework against Hybrid Cyber-Physical-Cognitive (CPC) Attacks

Alex Mathew

Bethany College, USA

***Abstract*— Modern hybrid campaigns coordinating the cyber, physical, and cognitive (CPC) attack vectors have been facilitated by offensive integration of artificial intelligence (AI) (Ostreni, 2021). These cross-domain operations exploit the seams between traditionally siloed defensive postures, undermining critical infrastructure and sociopolitical stability. The given paper presents SHIELD (Strategic Hybrid-incident Early-warning and Layered Defense) as a unified AI infrastructure that is capable of automatically identifying, correlating, and setting boundaries to the next generation CPC attacks. SHIELD contributes in three main technical ways, namely (1) Multimodal Context-Aware Deepfake Detection (M-CADD), which is a pipeline that integrates an active semantic claims-based and passive contextual fact-verification probing against repositories of trusted knowledge to detect linguistic anomalies in synthetic media, with an accuracy rate of 94.3% due to simulated influence operations; (2) Resilient Control System Immune Learning (RCS-IL), a reinforcement learning (RL) agent trained in a high-fidelity digital twin environment to maintain operational stability under stealthy, multi-stage attacks, reducing impact severity by 78% compared to signature-based defenses; and (3) Cross-Domain Hybrid Attack Graph Engine (C-HAGE), a probabilistic graphical model that fuses sociotechnical, cyber, and physical sensor data to reconstruct adversarial kill chains, predicting next-stage attack actions with 82% precision. The effectiveness of the framework is supported by simulation on industrial control systems (ICS) and disinformation testbeds, demonstrating that fused, cross-domain AI is critical for mitigating the compounded risks of hybrid threats.**

***Keywords*— Adversarial Machine Learning, Cyber-Physical Systems (CPS), Disinformation, Multimodal Fusion, Resilient Control, Cross-Domain Threat Intelligence.**

I. INTRODUCTION

The antagonistic implementation of AI has signaled a paradigm shift in threat models, which has shifted towards multi-domain intrusions to Hybrid Cyber-Physical-Cognitive (CPC) campaigns (Chesney & Citron, 2023). The synergistic effect on these operations is such that cognitive disinformation compromises the organization's vigilance in order to facilitate accuracy in cyber-physical exploitation (Sholademi, 2024).

Defensive fragmentation—where security operations centers (SOCs), physical security teams, and information integrity units operate in isolation—creates exploitable seams.

The hypothesis of the current paper assumes that successful defense needs an equally incorporated AI structure. We introduce the SHIELD framework to respond to the overall study question: How can multimodal AI methods, when applied in fusion, autonomously correlate and alleviate tri-domain hybrid attacks? SHIELD helps add three new technical functions, focusing on critical points of the CPC model, to allow predictive threat intelligence and automated response.

II. RELATED WORK

• *Synthetic Media Detection:* The existing SOTA approaches use low-level artifact detection (e.g., spectral anomalies, inconsistent generative adversarial network footprint), which cannot resist diffusion-based models (Zheng et al., 2025; Lai et al., 2025) and neglect artificially congruent semantic-contextual inconsistencies (Chi et al., 2025; Li et al., 2025). Some of the recent solutions involve cross-modal communication and contextual fusion (He et al., 2025).

• *ICS Security:* Older methods, like protocol whitelisting and shallow anomaly detection, cannot work against false data injection (FDI) on AI-generated false data attacks, which honor system dynamics (Giraldo et al., 2023). Digital twins are typically applied to simulation, but the use of digital twins in training adaptive RL-based defense agents has been poorly studied (Bak et al., 2025; Tang et al., 2025).

• *Cross-Domain Threat Intelligence:* Attack graphs are useful in modelling network penetration of IT networks, but do not include sociotechnical and physical sensor data (Mittal et al., 2024; Qiu et al., 2024). Even though the method of heterogeneous graph neural network is shown to be beneficial in terms of both temporal intrusion detection (King et al., 2023) and analyzing APT campaigns (Bahar et al., 2025), it is yet to be applied to real-time CPC fusion (Wang et al., 2024).

III. SHIELD Architectural Review

SHIELD uses a multi-agent, federated architecture. There are three fundamental AI engines, and they are M-CADD, RCS-IL, and C-HAGE, which are used in their specific domains (cognitive, physical, cyber). An enterprise Fusion & Orchestration Controller (FOC) provides alert correlation and contextual, enriching services, and orchestrates mitigative activity across domains with a shared Hybrid Attack Ontology based on predefined playbooks. Such collaborative systems take structures that focus on privacy issues as structured in line with privacy-driven ML frameworks (Boar et al., 2025; Shahriar et al., 2023).

IV. Pillar I: Multimodal Context-Awarm Deepfake Detection (M-CADD).

Hypothesis: M-CADD assumes that the practical logical inconsistency is a higher measure of organized campaigns with malicious synthetic media than low-level artifacts alone (Guo et al., 2022).

A. Architecture & Methodology

1. Multimodal Input Processing: Ingests video V, audio A, metadata M, and real-time contextual feeds Ct (e.g., geolocation, calendar, financial, and environmental data).
2. Semantic Claim Extraction: A fine-tuned vision-language model (e.g., ViT-LLaMA) processes V and A to output a set of semantic claims \(S = \{s_1, s_2, \dots, s_n\}\)
3. Contextual Verification Engine: Each claim s_i is checked against a curated knowledge graph KG (Mahid et al., 2025) and real-time feeds Ct. A consistency score ψ_i ∈ [0, 1] is computed using transformer-based entailment models.
4. Dissemination Graph Analysis: Constructs a propagation graph Gd=(N, E) from metadata M. Extracts graph-theoretic features (e.g., burstiness, bot-cluster ratio) to yield an anomalous spread score α.
5. Fusion Classification: A feed-forward neural network F_θ takes the feature vector [ψ, α, φ(V, A)] (where φ denotes SOTA artifact features) to produce the final probability P(malicious).

B. Experimental Validation

We constructed a dataset of 500 high-quality deepfakes (generated using Stable Diffusion 3.0 and Wav2Lip 2.0) embedded within plausible false narratives. The prediction with high Baselines (SOTA: 81.2% accuracy) attempts were not as good as with m-cADD (M-CADD: 94.3% accuracy, F1-score: 0.927). It showed high levels of effectiveness in indicating logically impossible claims even at a high visual fidelity.

V. Pillar II: Resilient Control System (RCS) Immune Learning (RCS-IL)

RCS-IL formulates ICS defense as a Partially Observable Markov Decision Process (POMDP)—a modeling approach also effective for assessing system availability and security (Kharchenko et al., 2022)—which is solved via RL within a high-fidelity digital twin.

A. Digital Twin and Adversarial Environment.

Physical dynamics of the twin models (e.g., equations of power flow), the control logic, and network layers. A multi-vector (FDI + actuator compromise) adversarial RL agent (Chen et al., 2024) is conditioned to instigate a dynamic threat environment to the defender agent, which is similar to autonomous systems' physical testing (Wang et al., 2023).

B. RL Formulation

- State Space s<sub>t</sub>: It is the combination of sensor measurements y.t., actuator states u.t., and network warnings n.t., and the resultant physical invariants h (y.t., u.t.).
- Action Space a<sub>t</sub>: Discrete continuous mixture, e.g., {isolate substation, adjust setpoint, override PLC command}.
- Reward Function $R(s_t, a_t)$:
- $R = -(\lambda_1 \cdot \|\Delta f\|_2 + \lambda_2 \cdot L\_load + \lambda_3 \cdot C\_intervention) + \lambda_4 \cdot \mathbb{1}\_detection$
- where Δf is frequency deviation, L_load is load shed, and C_intervention is the cost of defensive actions.

C. Training & Results

In a simulated 72-hour attack on an IEEE 39-bus model, RCS-IL reduced the Impact Severity Index (ISI) by 78% compared to a rule-based intrusion detection system (IDS), while maintaining frequency within ±0.15 Hz—demonstrating robustness that could be further analyzed using reachability methods (Zhang et al., 2023).

VI. Pillar III: Cross-Domain Hybrid Attack Graph Engine (C-HAGE).

C-HAGE employs a Temporal Heterogeneous Graph Neural Network (THGNN) to model CPC attack progression, extending concepts from spatial-temporal graph models used in APT detection (Bahar et al., 2025)

A. Graph Construction

- *Nodes V:* Here will be instances that belong to domains (e.g., pipe, SocialMediaAccount, Firewall, PLC).
- *Edges E:* Indicate witnessed relationships or attack event (ex: POSTS, EXPLOITS, CONTROLS).
- *Node/Edge Attribution:* Each node and edge is attributed with a time-varying compromise probability $P_c(v_i, t)$. Bayesian belief update: Each node and edge is attributed with a time-varying compromise probability P.

B. Probabilistic Inference & Prediction

THGNN learns using old attack graphs. The likelihood of an attack step e_{jk}^{t+1} at time *t + 1*is:

$$P(e_{jk}^{(t+1)}) = \sigma(\mathrm{THGNN}(h_j^{(t)}, h_k^{(t)}, P_c^{(t)}, e_{hist}))$$

Where:
$h_j^{(t)}$ = embedding of node j at time t
$P_c^{(t)}$ = vector of compromise probabilities at time t
e_{hist} = historical edge features
σ = sigmoid activation function

C. Evaluation

In a simulated smart-city attack hybrid attack, C-HAGE was 82 percent accurate in the target stage of the next attack (e.g., social media boom to targeted malware deployment of a SCADA set), with an average lead time of 23 minutes at which countermeasures could be taken.

VII. Integrated Case Study: Election Infrastructure Defense).

The integrated response of SHIELD was tested with the help of a simulated campaign against election infrastructure:

1. *Cognitive:* On the geolocation of videos and schedule mismatch, M-CADD spotted the occurrence of a deepfake video of an official as contextually inconsistent (P = 0.96).
2. *Correlation:* C-HAGE performed a correlation between the spread of the deepfake and simultaneous DDoS warnings, increasing the level of danger and making the possibility of GPS spoofing of logistics systems probable.
3. *Physical Mitigation:* Physical Mitigation RCS-IL (modified to suit fleet management) did not recognize the spoofed GPS signal and activated a congestion to inertial navigation to avoid disruption.

The FOC orchestrated a unified response: issuing public rebuttals, redirecting DDoS traffic, and securing physical assets

VIII. Discussion, Limitations, And Future Work.

Discussion: SHIELD proves that the synergy of hybrid attacks can be disturbed by cross-domain AI fusion. Predictive correlation's main strength is the capability to extrapolate latent correlations across domains of CPC.

Limitations:

- *Data Privacy:* M-CADD and C-HAGE will have access to sensitive data. To implement it in practice, federated learning and a differential privacy method should be combined (Li et al., 2024; Olowononi et al., 2021).
- *Adversarial Adaptation:* Adversaries can also introduce model extraction or poisoning attacks to SHIELD elements (Guo et al., 2022), which require adversarial training to be done continuously.
- *Computational Overhead:* Real-time inference, especially for the THGNN and digital twin, demands significant computational resources.

Future Work

Future directions consist of: 1) creating a Hybrid Threat LLM to Aid C-HAGE in the creation of threat reports; 2) implementing a hardware-in-the-loop testbed for real-world validation; and 3) formalizing the FOC's decision-making as a Stackelberg game against strategic adversaries.

IX. Conclusion

The growth of AI-driven hybrid warfare requires a paradigm shift from being domain-sized in defenses to autonomous and integrated AI systems. SHIELD provides a validated technical blueprint through its three core components: M-CADD for cognitive integrity, RCS-IL for physical resilience, and C-HAGE for cross-domain threat intelligence. Through a combination of AI in the cyber, physical, and cognitive planes, we will have the ability to protect the most vital nexus where these vectors meet. This combination is not just scholarly but a practical need to protect critical infrastructure and other systems of society.

REFERENCES

[1] Bahar, A. A. M., Ferrahi, K. S., Messai, M. L., Seba, H., & Amrouche, K. (2025). CONTINUUM: Detecting APT Attacks through Spatial-Temporal Graph Neural Networks. arXiv preprint arXiv:2501.02981. https://arxiv.org/pdf/2501.02981

[2] Boar, P. E., Sharma, P. K., Akram, R. N., & Li, W. (2025). SoK: Towards Privacy-Centric Collaborative Machine Learning—A Classification Framework for Privacy Solutions. SN Computer Science, 6(8), 972. https://link.springer.com/article/10.1007/s42979-025-04482-4

[3] Chi, Z., Guo, P., & Liu, F. (2025). A Compact GPT-Based Multimodal Fake News Detection Model with Context-Aware Fusion. Electronics, 14(23), 4755. https://doi.org/10.3390/electronics14234755

[4] Guo, W., Tondi, B., & Barni, M. (2022). An overview of backdoor attacks against deep neural networks and possible defences. IEEE Open Journal of Signal Processing, 3, 261-287. https://ieeexplore.ieee.org/abstract/document/9827581/

[5] He, J., Liu, T., Zhao, J., & Turner, B. (2025). MM-FusionNet: Context-Aware Dynamic Fusion for Multi-modal Fake News Detection with Large Vision-Language Models. arXiv preprint arXiv:2508.05684. https://arxiv.org/pdf/2508.05684

[6] Kharchenko, V., Ponochovnyi, Y., Ivanchenko, O., Fesenko, H., & Illiashenko, O. (2022). Combining Markov and Semi-Markov Modelling for Assessing Availability and Cybersecurity of Cloud and IoT Systems. Cryptography, 6(3), 44. https://doi.org/10.3390/cryptography6030044

[7] King, I. J., Shu, X., Jang, J., Eykholt, K., Lee, T., & Huang, H. H. (2023, October). EdgeTorrent: Real-time temporal graph representations for intrusion detection. In Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (pp. 77-91). https://dl.acm.org/doi/epdf/10.1145/3607199.3607201

[8] Lai, B., Wang, X., Rambhatla, S., Rehg, J. M., Kira, Z., Girdhar, R., & Misra, I. (2025). Toward Diffusible High-Dimensional Latent Spaces: A Frequency Perspective. arXiv preprint arXiv:2511.22249. https://arxiv.org/pdf/2511.22249

[9] Li, X., Qiao, J., Yin, S., Wu, L., Gao, C., Wang, Z., & Li, X. (2025). A survey of multimodal fake news detection: a cross-modal interaction perspective. IEEE Transactions on Emerging Topics in Computational Intelligence. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10959073

[10] Mahid, Z. I. B., Manickam, S., & Kusuma, R. S. (2025). Fake News Detection: Intelligent Fact-Checking Based on Relational Similarity and Confidence Score in Knowledge Graph. Yayasan Ghalih Pelopor Pendidikan (Ghalih Foundation). https://books.google.com/books?hl=en&lr=&id=2e54EQAAQBAJ&oi=fnd&pg=PA1&dq=Fact-checking+the+frame:+A+knowledge-graph+approach+to+deepfake+mitigation.&ots=zODBRrB-PK&sig=iU2-f8Pqnx4cI0NfZDHJP2AaxrI

[11] Noorizadeh, M., Shakerpour, M., Meskin, N., Unal, D., & Khorasani, K. (2021). A Cyber-Security Methodology for a Cyber-Physical Industrial Control System Testbed. IEEE Access, 9, 16239–16253. https://doi.org/10.1109/access.2021.3053135

[12] Olowononi, F. O., Rawat, D. B., & Liu, C. (2021, January). Federated learning with differential privacy for resilient vehicular cyber physical systems. In 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC) (pp. 1-5). IEEE. https://ieeexplore.ieee.org/abstract/document/9369480/?casa_token=-IYWnVuTflYAAAAA:3FOIp2GyUXJfIu6-jc2L0fPW4vsHfAjYCuhXkrGHclko5Xc634LUk2XCfv-WBgBYXHQHXoDVV0fxssM

[13] Ostreni, B. (2021). Understanding Hybrid Warfare Constructivism and Ontological (in) Security. https://dspace.cuni.cz/handle/20.500.11956/152237

[14] Qiu, S., Shao, Z., Wang, J., Xu, S., & Fei, J. (2024). Research on Power Cyber-Physical Cross-Domain Attack Paths Based on Graph Knowledge. Applied Sciences, 14(14), 6189. https://doi.org/10.3390/app14146189

[15] Shahriar, S., Allana, S., Hazratifard, S. M., & Dara, R. (2023). A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. IEEE Access, 11, 61829–61854. https://doi.org/10.1109/ACCESS.2023.3287195

[16] [Sholademi, D. B. (2024). Leveraging AI for Detecting Deep Fakes and Combating Financial Fraudulent Identity Schemes. International Journal of Research Publication and Reviews, 5(12), 4096–4111. https://doi.org/10.55248/gengpi.5.1224.250131

[17] Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., & Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. Annual Review of Control, Robotics, and Autonomous Systems, 8(1), 153-188. https://www.annualreviews.org/content/journals/10.1146/annurev-control-030323-022510

[18] Urbinati, A. (2023). Analysis of sociotechnical systems: from data to complex networks models. https://tesidottorato.depositolegale.it/bitstream/20.500.14242/198261/1/Urbinati_Tesi.pdf

[19] Wang, N., Luo, Y., Sato, T., Xu, K., & Chen, Q. A. (2023). Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4412-4423). https://openaccess.thecvf.com/content/ICCV2023/papers/Wang_Does_Physical_Adversarial_Example_Really_Matter_to_Autonomous_Driving_Towards_ICCV_2023_paper.pdf

[20] Zhang, C., Ruan, W., & Xu, P. (2023, June). Reachability analysis of neural network control systems. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 12, pp. 15287-15295). https://ojs.aaai.org/index.php/AAAI/article/view/26783