American Journal of Interdisciplinary Research and Innovation (AJIRI)

ISSN: 2833-2237 (ONLINE

VOLUME 5 ISSUE 3 (2026)

PUBLISHED BY

E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Interdisciplinary Research and Innovation (AJIRI)

Volume 5 Issue 3, Year 2026
ISSN: 2833-2237 (Online)
DOI: https://doi.org/10.54536/ajiri.v5i3.8282
https://journals.e-palli.com/home/index.php/ajiri

Temporal Drift Modeling in Multimodal Frameworks for Sockpuppet Detection in Social Networks: A Theoretical Framework

Mohammed Mhothasim Ali[1*]

Article Information

Received: May 12, 2026

Accepted: July 25, 2026

Published: August 28, 2026

Keywords

Late Fusion, Multimodal Ai, Privacy-Preserving Ai, Social Network Security, Sockpuppet Detection, Temporal Drift

ABSTRACT

Sockpuppet accounts, multiple fake profiles controlled by a single operator, threaten the integrity of online social networks through coordinated disinformation and opinion manipulation. Existing detection methods rely on static snapshots of behavioral or textual features, leaving them blind to a defining sockpuppet characteristic: deliberate dormancy followed by coordinated reactivation. This paper proposes a four-modality late-fusion framework that extends the multimodal deduplication architecture into the social network domain. Three existing modalities, semantic text embeddings, static behavioral features, and device metadata, are adapted for the social network context. A fourth modality, Temporal Drift, is formally introduced, defined through the Activity Window, Baseline Profile, Drift Vector, Temporal Drift Score (TDS), Dormancy Threshold, and Burst Index. Grounded in time-series anomaly theory, the framework detects dormancy-to-activation behavioral signatures without relying on any personally identifiable information (PII). Building on an expanded review of the 2022–2026 literature on coordinated inauthentic behavior, sybil detection, concept drift, and trustworthy AI, a theoretical comparison with four prior works confirms the framework's unique position as, to the authors' knowledge, the only approach combining multimodal fusion, temporal drift sensitivity, privacy preservation, social network applicability, and unsupervised operation. Beyond cybersecurity, the framework contributes to trustworthy AI, privacy-preserving behavioral analytics, and interdisciplinary digital identity research spanning social computing, human–computer interaction, and digital governance.

INTRODUCTION

Indeed, coordinated inauthentic activity (such as sock puppetry, astroturfing and other forms of manufactured consensus) is a key way that misinformation, coordinated influence operations and online manipulation undermine trust in digital platforms. Sockpuppet accounts are a key element in these larger campaigns and are the first step in wider efforts to protect online discourse, election integrity and platform safety.

To spread disinformation, to create consensus and to amplify a narrative across social platforms, sock puppet accounts are duplicate accounts operated by one person. (Alharbi *et al.*, 2021) Sockpuppets are not automated bots, but are controlled by humans and adapt their behavior in an attempt to evade detection methods; they are unsuitable for detection with static or rules-based detection methods. Alharbi *et al.* (2021) present a survey on social media identity deception and make dormancy cycling the major open challenge.

Tsikerdekis and Zeadally (2014) have shown that it is possible to identify sockpuppetry on Wikipedia without using text analysis techniques. But their model scans for behavior in a fixed period in the early window, and accounts that don't occur in the early window are completely missed. Later, (Yamak *et al.*, 2018) suggested SocksCatch that clusters sockpuppets using machine learning and graph community detection methodologies, but considers all features as time-stable. Both do not model behavior over time, the one dimension that is used most intentionally by sophisticated operators.

(Ahmed, 2025); Tsikerdekis and Zeadally (2014) have shown that it is possible to identify sockpuppetry on Wikipedia without using text analysis techniques. But their model scans for behavior in a fixed period in the early window, and accounts that don't occur in the early window are completely missed. Later, (Yamak *et al.*, 2018) suggested SocksCatch that clusters sockpuppets using machine learning and graph community detection methodologies, but considers all features as time-stable. Both do not model behavior over time, the one dimension that is used most intentionally by sophisticated operators. Also built a multimodal deduplication system that preserves privacy of the data in CRM and healthcare, by combining semantic embeddings, login-cadence features, and device metadata through a late-fusion (as opposed to early-fusion, where each modality's features are combined) approach with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering. This system is privacy-oblivious, and can have a high recall in noisy, privacy-constrained data. The architecture is easily translatable to the social network application, as there are still the same three signals, albeit with the addition of a

[1] New York University, United States
[*] Corresponding author's e-mail: mohammedmhothasimali@pingmx.com

fourth: time.

This paper fills this void by formalizing and extending (Ahmed, 2025) by adding a Temporal Drift modality that is based on time-series anomaly theory (Zamanzadeh Darban *et al.*, 2024) and (Lazebnik & Iny, 2024) and adapting the full framework for duplicate profile detection in social networks. The contribution is theoretical: a formally defined, privacy-preserving, four-modality architecture that is identified as future work to be empirically validated. However, as described in Section II the literature on coordinated inauthentic behaviour, sybil detection and concept drift has continued to expand over the last few years, but in the absence of this work closing this specific gap, it is even more motivating for the present contribution. The contribution, instead of adding a single feature to an existing system, is more broadly a theoretical AI architecture, a formal model of temporal behavioral drift, a privacy-aware approach to multimodal learning, and an adaptable foundation for interdisciplinary digital trust research.

The proposed framework is not only in the field of computer science, but but it is a hybrid between various disciplines. It leverages Artificial Intelligence for the multimodal learning architecture, Social Computing and Human Behaviour Analytics for the model of dormancy-to-activation behavioural signatures, Cybersecurity for the adversarial threat model and Privacy Engineering for the explicit avoidance of personally identifiable information. The placement of the contribution aligns with the overall vision of this work – to identify a technical artefact, while also facilitating the trustworthy, privacy-respecting conduct of digital identity research across the spectrum of AI, behavioural science, and digital governance. Comparable AI architectures in critical-infrastructure research likewise combine predictive analytics, system models, and automated decision support, including digital-twin grid resilience, physics-informed grid twins, AI-driven grid control, and drone-based asset monitoring (Sunkara & Sunkara, 2026; Sunkara, 2026a; Sunkara, 2026b; Sunkara, 2026c). These examples are cited only as cross-domain architectural parallels, not as evidence for sockpuppet detection. A comparable cross-domain parallel is found in shared-spectrum networking, where coordinated multi-tier access and privacy-aware coexistence are treated as joint design constraints rather than separate concerns (Agarwal *et al.*, 2022; Manekiya *et al.*, 2020). This study focuses on these research questions.

1. How can temporal drift in user behavior, captured through constructs such as the Activity Window, Baseline Profile, Drift Vector, Temporal Drift Score, Dormancy Threshold, and Burst Index be formally modeled and integrated as a distinct modality within a multimodal late-fusion framework for sockpuppet detection?
2. Can a late-fusion architecture adapted from healthcare data deduplication achieve privacy-preserving sockpuppet and duplicate-profile detection in social networks without relying on personally identifiable information at any stage?
3. How does incorporating a temporal drift modality improve theoretical detection of dormancy-to-activation behavioral signatures compared to existing static, snapshot-based detection approaches, and can the resulting framework serve as a reusable foundation for future empirical validation?

LITERATURE REVIEW

Sockpuppet, Sybil, and Identity Deception Detection

Sockpuppet, Sybil, and Identity Deception Detection

In an independent study, Tsikerdekis & Zeadally 2014) showed that nonverbal behavioural signals, namely the timing and distribution of namespaces, can be used to identify sockpuppetry with high accuracy and are also very efficient to compute, without using text content. However, a fixed window design requires that the behaviour be stationary. This was later extended by (Yamak *et al.*, 2018), who achieved detection rates of 89-95% and grouping accuracy of 80-88% on Wikipedia data, but with the same limitation of static features. (Alharbi *et al.*, 2021) Review the more than 10 years of history of identity deception research and find that temporal evasion is still an unsolved problem in all dominant detection paradigms. Comparable behavioural feature-engineering strategies have proven effective in adjacent adversarial detection settings: Khaliq *et al.* (2023) show that behavioural attributes extracted from binaries and refined through feature engineering allow standard classifiers to identify backdoor malware that conceals itself through registry-key and API-call manipulation, reinforcing the broader principle that evasive adversaries are exposed more reliably by engineered behavioural signals than by static rules.

Structural rather than purely behavioral signals have been recently used to solve the related problem of detecting sybils and fake-accounts. Contrary to the common assumption in earlier belief-propagation methods, Loopy Belief Propagation (LBP) fails to take account of the local heterogeneity in a directed social graph, leading to inaccurate detection. To solve this problem, Lu *et al.* (2023) present a novel approach called SybilHP to estimate homophily at each node in the graph adaptively instead of making a single assumption, thereby improving the detection performance. Heeb and Plesner (2024) tackle the same problem using graph attention networks and demonstrate that learned attention weights over structural neighborhoods are superior to classical belief propagation baselines, especially when the attack-edge density is high. Breuer *et al.* (2023) take it one step further and suggest a preferential-attachment classifier to identify potential fake accounts early on in the process, before any account relationships are even formed, by monitoring unanswered friend requests. To complement these structural methods, Goyal *et al.* (2023) show that multimodal deep learning (fusing textual, behavioral, and profile data) outperforms single-modality baselines for fake-account detection, which aligns well with the late-fusion design of this framework. Similarly, Khan *et al.* (2024) present work on graph-based reasoning beyond mere duplication, using graph-neural networks

trained with data from social media platforms to detect malicious users. Similarly, Moore (2023) provides a non-technical view, suggesting that the fact that this epistemic uncertainty about the number of fake accounts on major platforms persists makes it even more imperative to have independent and auditable detection methods; as a result, this paper focuses on specifying formally and inspectable constructs, as opposed to learned representations that are opaque.

Multimodal Fusion and Coordinated-Account Detection Architectures

For multimodal detection, (Li, 2024) present a transformer-based bot detection method that combines text, user statistics, and metadata from social network information; they claim that multimodal signal fusion greatly surpasses the single-modal baselines. The present system rests on the three-modality late-fusion proposed by (Ahmed, 2025) which achieved an F1-score of 0.665 and recall of 0.995 on the CRM deduplication without PII task. The semantic modality design is confirmed by (Paganelli *et al.*, 2022) via systematic benchmarking, where BERT-based embeddings outperform. The advantage of combining heterogeneous sensing streams is not confined to social platforms: Saha *et al.* (2025) review deep-learning applications over multi-sensor Earth observation data and report that fusing modalities with differing resolutions and noise characteristics consistently outperforms single-sensor pipelines, an argument structurally parallel to the multimodal case made here. Cross-domain engineering studies also illustrate AI pipelines that couple learned representations with real-time monitoring and decision support (Sunkara, 2026a; Sunkara, 2026b; Sunkara, 2026c). Their relevance here is methodological rather than domain-specific.

Much work is being done that looks at coordinated, rather than individual, account behavior. Detection methods take a step beyond single-account features to be effective, as shown by Coordinated Inauthentic Activity on Twitter and its measurable impact on information-sharing dynamics in Cinelli *et al.* (2022). This approach is conceptually close to the Burst Index construct used in this framework, but it is used for account creation time, instead of reactivation time, as done by Bellutta and Carley (2023) in their work on detecting coordinated account creation. In this consolidating survey, Mannocci *et al.* (2024) analyse the current literature in the field of coordination-detection by grouping methods according to the actors, actions and intent being analysed; Mannocci (2025) further extends this consolidation into a dedicated methodological framework for the analysis of coordinated behaviour. Cinus *et al.* (2024) present cross-platform coordinated-activity detection at the 2024 U.S. election, and (Luceri *et al.*, 2025) add video-first platforms like TikTok to the mix, revealing that beyond text-based platforms, multimodal signals – visual, textual, and behavioural – are needed for coordination detection. (Bellutta & Carley, 2023)further nuance the picture by demonstrating empirically that much of coordinated link sharing on Facebook is not necessarily 'malicious' but is clearly motivated by commerce or organizations, and thus the need for a framework such as the one proposed here, which does not conflate coordinated link sharing with inauthenticity. This paper has focused on the unsupervised clustering approach instead of the labeled classification approach, as noted by Cima *et al.* (2024) and Zouzou and Varol (2024) who reported unsupervised methods for detecting coordinated information operations and fake-follower campaigns respectively. An empirical example of an emerging generation of AI-generated "sleeper" social bot that can be dormant for long periods of time, and then coordinated to strike, is described by (Doshi *et al.*, 2024) have introduced an unsupervised social-bot detector based on structural information theory, while Aljabri *et al.* (2023) have compiled a thorough review that verifies that the detection of social bots by machine learning is an active and diverse area where there is no dominant paradigm.

Temporal Drift, Concept Drift, and Anomaly Detection

To address the issue of temporal grounding, Zamanzadeh Darban *et al.* (2024) offer a comprehensive taxonomy of deep learning anomaly detection for the subsequent anomaly class: patterns anomalous only as sequences over time; and (Lazebnik & Iny, 2024) show how to apply drift-based anomaly detection to social media interaction graphs.

The Temporal Drift modality might have some additional theoretical support from the wider concept-drift literature. Xiang *et al.* (2023) provide an overview of deep-learning-based concept-drift adaptation methods, which are divided into two categories: gradual concept drift and abrupt concept drift, and list some methods that can be applied to the Drift Vector formulation proposed in this paper. Ding *et al.* (2022) present a transformer-based concept-drift adaptation method for time-series anomaly detection and conclude that the models explicitly aware of concept drift outperform stationary-based models on benchmarks that have non-stationary behavior, supporting the main claim of this paper that stationary-based models are structurally disadvantaged in the presence of concept drift. Liu *et al.* (2023) consider a closely related problem: joint detection of concept drift and anomalies in time series, in which they need to separate the Dormancy Threshold (a change point) from the Temporal Drift Score (an anomaly magnitude), which is similar to this work's requirement to distinguish the Dormancy Threshold from the Temporal Drift Score. Zhu *et al.* (2023) present METER, a dynamic notion of "normal" to incorporate online anomaly detection, which continuously updates the definition of "normal" behavior, and they test it on an online crime data set.They propose an online crime dataset and test the METER framework on the dataset by Zhu *et al.* (2023), which is a dynamic concept-adaptation framework for online

anomaly detection that continuously updates its notion of "normal" behavior. Providing an alternative design to this framework's fixed Baseline Profile that can be investigated for long-lived accounts in future work. It is also an open and active research topic, one that is not exclusive to the social network area, as Li *et al.* (2023) confirm: "The challenge of keeping a stable baseline while accounting for a changing one is an open and highly studied problem.

Trustworthy, Explainable, and Privacy-Preserving AI for Platform Governance

In addition to the immediate detection literature the framework also builds on interdisciplinary research on trustworthy and explainable AI. Trustworthy AI systems have technical and organizational requirements as discussed by Kaur *et al.* (Kaur *et al.*, 2022) and (Thiebes *et al.*, 2020) and explainable AI techniques applicable to auditing AI automated detection decisions are surveyed by Barredo Arrieta *et al.* (Barredo Arrieta *et al.*, 2020) and Chamola *et al.* (Chamola *et al.*, 2023). From a social and platform-governance perspective, (Gruzd *et al.*, 2023) discuss the relationship between content moderation and misinformation, and the broader subject of digital-trust and platform-governance, placing this detection effort in the context of other efforts.

What we see in more recent studies on explainability clarifies these images. Saeed and Omlin (2023) give a systematic meta-survey of explainable AI, enumerating longstanding challenges that any new audit layer for this framework would have to face, such as the problem of evaluating the quality of explanations themselves. Both Yang *et al.* (2023) and Yang *et al.* (2023) have also surveyed the approaches to XAI and their limitations, and in a context of fraud detection similar to that in this paper, Zhong and Goel (2024) showed that incorporating an explainability layer into a black box classifier positively impacted auditor trust and transparency in decision making. The work presented here complements the categorical and aggregation-based privacy model used here, as there is a complementary formal model in the literature, in this case proposed by Jiang *et al.* (2023) for differential-privacy applications in social network analysis, and other work, in this case by Fu *et al.* (2022), for privacy-preserving graph analytics using secure generation and federated learning, which directly supports the proposed fifth modality, graph-structural (Section X). Combined, this literature puts the present framework's privacy-by-design position and its request for future explainability into an active and quickly developing research agenda, not as new ideas on their own.

Problem Formalization

Two accounts u_i and u_j are called duplicates if and only if $o(u_i) = o(u_j)$ and $i \neq j$. It is the detection task to find out all such pairs, while there is no access to any PII.

Dynamic behavioral models are not effective against adaptive sockpuppets, because they rely on the assumption of behavioral stationarity: the feature vector calculated on an initial observation window should be sufficient to characterize the current behaviour. Sophisticated operators deliberately make use of this assumption by masking their activities as normal in every observation window, and as coordinated activity in dormancy windows between observations. Anomaly taxonomy by Zamanzadeh Darban *et al.* (2024) is a subsequent anomaly with no single timestep, although the pattern of silent to coordinated activation is very discriminative when compared to an account's own behavioral history.

Once a problem has been well defined and the behaviour to be detected after the attack has been defined, the paper now moves on to the solution; that is, Section IV lays out the proposed multimodal framework, Section V describes the methodology used for the generation and evaluation of the framework and Section VI presents the theoretical analysis.

Table 1: Notation Summary

Symbol	Meaning
U = (Tsikerdekis & Zeadally, 2014)	Set of accounts on the social network
$o(u_i)$	Real-world operator of account u_i
W_t	Activity Window: sliding time slice of duration τ ending at t
b_t	Behavioral feature vector produced by window W_t
B	Baseline Profile: mean behavioral vector over the earliest k windows
D(t)	Drift Vector: signed deviation $b_t - B$
TDS(t)	Temporal Drift Score: L2 norm of the Drift Vector, $\|\|D(t)\|\|_2$
δ	Dormancy Threshold: minimum consecutive zero-activity windows
β	Burst Index: post-dormancy activity magnitude ratio

MTERIALS AND METHODS

Research Design

This paper is theoretical in nature and falls under the paradigm of design-science research (DSR) which is not empirical in nature. In cases where the most important contribution is a new artifact, in this case a formally specified detection architecture, and the value of the artifact can only be determined by careful conceptual design and comparative analysis, prior to empirical

instantiation, design-science research is well suited. It goes through four steps: (1) systematic review of previous work to find an unfilled capability gap; (2) formal mathematical specification of new constructs required to

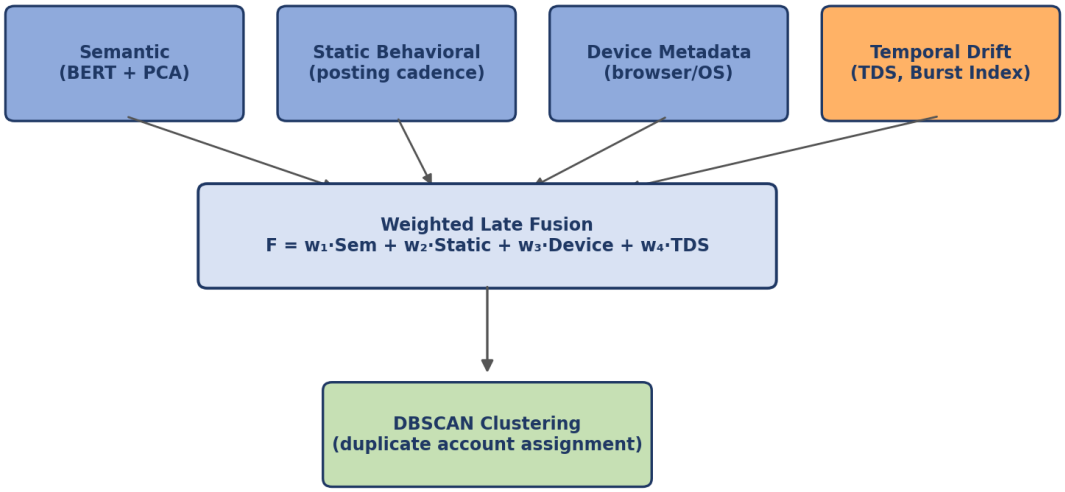

Figure 1: Inherited Modalities (Adapted from (Ahmed, 2025)

fill the gap; (3) architectural integration of new constructs with an existing system that has been extensively tested in practice; and (4) structured comparison of the resulting architecture to representative previous systems, using explicit criteria.

Architecture Overview

The proposed framework is built on top of (Ahmed, 2025) to explicitly model temporal behavioural drift in a fourth modality beyond the original three. All streams are independent and run concurrently, and they are fed to a weighted late-fusion layer, which is then clustered by DBSCAN. The use of late fusion is better than early fusion due to the fact that the modalities are structurally different. Li (2024) Also, the cross-modal normalisation distortions are avoided when they are combined at the decision level, as well as the graceful degradation. This is depicted in Fig. 1. Architecture-level emphasis on predictive digital representations and automated corrective action is also visible in recent grid-resilience and physics-informed power-flow work (Sunkara & Sunkara, 2026; Sunkara, 2026b; Sunkara, 2026c). This comparison concerns system-design principles rather than application equivalence. Evidence from multi-sensor fusion pipelines points in the same direction, indicating that heterogeneous streams are combined most reliably at the decision level once each stream has been characterised independently (Saha *et al.*, 2025).

Semantic Modality: Pre-trained transformer embeddings (BERT) of profile bio content and post content, reduced by PCA. The work of justifies the design (Paganelli *et al.*, 2022) who confirm that BERT is very good at unsupervised entity matching. Bio text and writing style are similar to the name/city field in the CRM context in terms of disambiguation (Ahmed, 2025).

Static Behavioral Modality: Statistics of posting behaviour, posting frequency; time-of-day distribution; inter-posting time; ratio of posting replies. The login-cadence modality is adapted directly from (Ahmed, 2025) replacing login timestamps with posting timestamps.

The Device Metadata Modality is the label-encoded browser, OS and access method (mobile API vs. web). (Li, 2024)Static Behavioral Modality: Statistics of posting behaviour, posting frequency; time-of-day distribution; inter-posting time; ratio of posting replies. The login-cadence modality is adapted directly from (Ahmed, 2025) replacing login timestamps with posting timestamps. The discriminative value of this modality depends on the heterogeneity of the access technologies available to users, which continues to widen as cellular networks extend into unlicensed and shared spectrum (Manekiya *et al.*, 2020). Shared-access frameworks such as the Citizens Broadband Radio Service further diversify the pool of access technologies and device classes that a platform may observe (Agarwal *et al.*, 2022).

Confirm that contributions from device-level signal does improve detection accuracy in the social media domain.

Temporal Drift Modality (Novel Contribution)

The Temporal Drift modality is a formal representation of the change in behavior over time. There are 6 constructs: Activity Window W: A time window of length τ that is applied to the posting history of an account. Window W_t of the interval $[t - \tau, t]$ yields a behavioral feature vector b_t.

Baseline Profile B: Mean of the behavioral vector over the first k windows of the account: $B = (1/k)\ \Sigma_{i=1}^{k}\ b_i$. Computed only once and stored as the reference against which to check for normal behaviour.

Drift Vector D(t): The signed difference between

the current window and baseline: $D(t) = b_t - B$. Each dimension represents the directionality of one behavioral dimension.

Temporal Drift Score TDS(t): L2 norm of the Drift Vector: $TDS(t) = ||D(t)||_2$. A scalar anomaly signal – high levels mean that the account is substantially different from its own history.

Dormancy Threshold δ: Minimum number of consecutive windows with zero activity to be considered a dormancy period. Platform-specific and conforms to the norm of inactivity of a normal user.

The Burst Index β is the ratio of overall activity magnitude in the first window after dormancy to the mean activity magnitude prior to dormancy, that is, $\beta = ||b_t||_2 / \text{mean}(||b_1||_2,...,||b_n||_2)$ with the norm $||\bullet||_2$ being the L2 norm of each behavioral feature vector, which yields a well-defined scalar value instead of a vector ratio. The most important sockpuppet activation signature is a high β after a dormancy period that is greater than δ.

Such constructs implement the following anomaly class of . (Zamanzadeh Darban *et al.*, 2024) for user behavior, and correspond to the temporal drift benchmarks in social media of (Lazebnik & Iny, 2024). Further, they are consistent with the concept-drift adaptation literature discussed in Section II.Among the most significant contributions, Ding *et al.* (2022) provide a transformer that is drift-sensitive, and Liu *et al.* (2023) propose a joint anomaly/change-point formulation.

Late Fusion

The algorithm is: $F = w_1 \bullet \text{Semantic} + w_2 \bullet \text{StaticBehavior} + w_3 \bullet \text{Device} + w_4 \bullet \text{TDS}$

where $w_1 + w_2 + w_3 + w_4 = 1$. Weight w_4 scales with baseline stability: accounts with long-term history get higher w_4; short-term accounts get high w_2 until enough behavioral history is accumulated. The unsupervised clustering is done by clustering the fused vector F with DBSCAN, and the membership to the clusters becomes the duplicate assignment. There's no PII going through any modality at any time.

Literature Synthesis and Gap Identification

As in the Literature Review (Section II), four previous detection systems were chosen as comparators because of their direct relevance to sockpuppet or duplicate-account detection: the two most cited ones for the social network domain (Tsikerdekis & Zeadally, 2014) and (Yamak *et al.*, 2018) respectively, the representative multimodal bot-detection system of (Li, 2024)Literature Synthesis and Gap Identification and the direct architectural predecessor of the present framework, (Ahmed, 2025) Literature Synthesis and Gap Identification. They were each analysed in terms of the presence or absence of five properties, which were considered essential for a detector to be: Multimodal, Unsupervised, Temporally-aware, Privacy-preserving, and Applicable to the social network domain. This paper resolves the capability gap that existed with no system capable of meeting all the above properties.

Formal Construct Specification

This is the Temporal Drift modality (Section IV.C) was developed by making the general subsequent-anomaly formalism described in (Zamanzadeh Darban *et al.*, 2024) (where no individual timestep is anomalous, but a transition between behavioral regimes is) specific to account-level social media behavior. Each construct (Activity Window, Baseline Profile, Drift Vector, Temporal Drift Score, Dormancy Threshold and Burst Index) was designed to be computable from a set of non-identifying behavioral statistics only, following the privacy constraint adopted by (Ahmed, 2025).

Architectural Integration Method

The three inherited modalities (semantic, static behavioral, device metadata), which were adapted from Ahmed's (Ahmed, 2025) healthcare deduplication architecture, were changed by direct field-for-field substitution (e.g., replacing login timestamps with posting timestamps), and were not redesigned, as (Li, 2024) and (Paganelli *et al.*, 2022) independently found that the same signal types (text embeddings, behavioral cadence, and metadata) work well in the social network domain. Then, the Temporal Drift modality was added as a fourth, independent late-fusion stream, following the late-fusion argument of (Ahmed, 2025) that heterogeneous modalities should be combined only at the decision layer.

Comparative Evaluation Method

This four-modality architecture was subjected to qualitative performance comparison to the four comparator systems in terms of the five criteria outlined in Section V.B (summarized in Table I, Section VI.A). The absence of a shared benchmark or labelled dataset in all five systems makes the evaluation qualitative and criterion-based, as opposed to metric-based, and the unit

Table 2: Framework Comparison

Property	(Yamak *et al.*, 2018)	(Tsikerdekis & Zeadally, 2014)	(Li, 2024)	(Ahmed, 2025)	This Framework
Modalities	1	1	3	3	4
Temporal Drift	No	Partial	No	No	Yes
Privacy-Preserving	Partial	Yes	No	Yes	Yes
Social Context	Yes	Yes	Yes	No	Yes
Unsupervised	Yes	Yes	No	Yes	Yes

of comparison are the criteria themselves, rather than any numerical score (see Section X for an explicit and acknowledged limitation of the present methodology).

RESULTS AND DISCUSSION

Results of applying the methodology of Section V are three: (1) a structured comparison with previous systems, (2) a qualitative adversarial-robustness assessment and (3) a qualitative privacy assessment. As the contribution is theoretical, these results reflect the analytically-determined properties of the proposed architecture, rather than empirical performance measurements taken on a benchmark with labeled sockpuppet sets; the necessary empirical work is proposed to be primarily in the direction of quantitative benchmarking against labeled sockpuppet sets (Section X).

Comparative Positioning Result

Table 2 shows the proposed framework is the only one, among the five systems considered, that simultaneously satisfies all five criteria: four modalities, explicit temporal-drift sensitivity, full privacy preservation, applicability to the social network domain, and unsupervised operation. While satisfying the social-context and unsupervised criteria, (Yamak *et al.*, 2018) and (Tsikerdekis & Zeadally, 2014)are both restricted to a single modality and, at best, partial temporal sensitivity. (Li, 2024) provides multimodality but has no privacy or temporal drift support. (Ahmed, 2025) meets four out of the five requirements, but was not created for the social network space and features no temporal structure. In fact, to the authors' knowledge, based on the literature search which included the recent coordinated-behavior and sybil-detection systems of Bellutta and Carley (2023), Lu *et al.* (2023), Heeb and Plesner (2024), and Cinus *et al.* (2024), no published system combines all five properties, further strengthening the novelty claim.

Adversarial Robustness Result

If the operator can evade one modality (such as the posting timings or the device itself), he cannot avoid the four-modality detector, as he needs to be able to ensure consistency in all the modalities (semantic content, static behavioural patterns, device fingerprints, posting timings) across all the duplicate accounts. This framework's four-modality structure directly implements the findings on adversarial adaptation by(Alharbi *et al.*, 2021), Breuer *et al.* (2023), and Heeb and Plesner (2024): hardening a single detector is not the most effective general defence; increasing the cost of evasion across multiple independent signal dimensions is. The same principle is reported in behavioural malware analysis, where adversaries that conceal themselves at one observation layer remain detectable through engineered features drawn from another (Khaliq *et al.*, 2023). Device-level consistency is likewise costly to fabricate as access technologies proliferate across licensed, unlicensed, and shared bands (Manekiya *et al.*, 2020).

Privacy Preservation Result

All 4 modalities work on a non-PII signal: text is converted to abstract embeddings, timestamps are aggregated into statistical features, device data is categorized and encoded into a dimension, and drift constructs are derived metrics (as opposed to actual values). This places the framework in the context of differential-privacy formalism, which is often required to explicitly specify a privacy budget and accept some utility loss in return, as the one presented by Jiang *et al.* (2023); in the current framework, on the other hand, privacy could not be compromised from the start of feature design, and designed-in privacy is the utility loss that is explicitly recognised and delineated here, rather than left implicit; the privacy-preserving graph analytics techniques of Fu *et al.* (2022) might be able to formalize further the utility loss should a fifth, graph-structural modality be added in future work.

Discussion

This study has outlined the theoretical methodology to the detection of sock puppets in Section V, and the results of this in Section VI, but the purpose of this section is to explain why each of the resulting design decisions is important, not just for sockpuppet detection, but beyond. Temporal drift is important because it attacks the specific behavioural approach that makes it hard to catch the more sophisticated sockpuppets: deliberate dormancy. Static detection methods make an implicit assumption that the behavior observed within a small-time window is representative of the behavior observed over time, an assumption that operators exploit by not saying a word during the observation window and only coordinating activity once attention is drawn elsewhere. The Temporal Drift modality represents a new way to consider the problem of detection by using the trajectory of the problem, rather than a snapshot of the problem. This effect is not limited to sockpuppet detection, either: the ability to time behaviors around observation windows (coordinated bot activation, review-bombing, or bought-engagement campaigns) could be formalized similarly and would be plausibly applicable to other behavioral-integrity issues in which adversaries can time their behaviors around observation windows.

The importance of late fusion is to keep the independence of heterogeneous types of signals. The statistical structure of these signals is very different, with the semantic signal typically structured at a higher level than the behavioral or device signals or the temporal signal, and forcing them into a common representation too early in the pipeline can lead to one modality's noise adding to the signal for another modality. This framework only integrates at the decision layer has two practical advantages: first, it allows the loss of a modality to be tolerated because it causes the system to degrade gracefully if the device metadata is not available; second, it ensures each modality's contribution can be inspected and adjusted independently as needed for auditing purposes and for the adaptive weighting

scheme future empirical work will require to calibrate. Comparable reasoning underpins resource-allocation work in unlicensed and shared bands, where independent decision layers are preferred to tightly coupled optimisation because operating conditions differ across users, technologies, and time (Manekiya *et al.*, 2020).

The reason privacy is important is that it's often the populations most susceptible to the harms of surveillance that are targeted by behavioral detection systems, particularly activists, journalists, and "ordinary" users who fall into "dragnet moderation." In this framework, by construction, all modalities operate on abstracted, aggregated, or categorically encoded signals, not on raw identifying content, thus maintaining the architecture's usability in regulated environments and minimising (but not eliminating) the risk that surveillance, rather than integrity protection, is achieved during deployment. However, architectural privacy does not mean a privacy-respecting deployment, as a governance issue, how a platform will use the detection outputs is not necessarily a straightforward technical question. Analogous tensions between coordinated multi-tier access and privacy preservation have been documented as governance problems, rather than purely technical ones, in shared-spectrum systems (Agarwal *et al.*, 2022).

These design considerations are indicative of design implications that go beyond the direct engineering contribution. In theory, the framework provides a blueprint for incorporating temporal, privacy-preserving reasoning into existing multimodal detection systems without redesigning them from scratch, as with Temporal Drift, which was added as a completely separate modality to Ahmed's three-modality system (Ahmed, 2025) rather than a complete rewrite. In sum, the same architectural structure – independent modalities that are fused at the end, and with an explicit temporal element – could plausibly be employed in adjacent trust-and-safety issues other than sockpuppet detection, as is explored in Section VIII. From a methodological point-of-view, the model's exclusive use of formally defined and mathematically specified constructs, in contrast to opaque learned representations, should allow it to be directly benchmarked when empirical validation is considered, and, at the same time, provide a stable point of reference for future researchers to compare it to other alternative temporal formulations, including the drift-adaptive models of Zhu *et al.* (2023) and Li *et al.* (2023). The transferability argument also has a methodological parallel in AI-enabled condition-monitoring and physics-informed operational workflows in critical infrastructure (Sunkara, 2026a; Sunkara, 2026c). These studies do not validate sockpuppet-detection outcomes.

All these implications are not empirically confirmed in this paper; they are simply the justifications for the architectural decisions made, not proven results. The other sections shift from an internal log to an external one, focusing on the framework's relevance to society, its potential for external application, and its ethical consequences, and end with the limitations that constrain all the claims made in this work.

Societal Significance and Practical Implications

Societal Significance

While the framework is described as a technical contribution, the challenge of sockpuppets is underpinned by several societal issues which are relevant beyond cybersecurity. Fake account coordination is an observed method for misinformation and manufactured consensus, and dormancy timings are specifically linked to attempts to impact public discourse on elections and other high-salience events. The earlier and more accurate detection of these accounts is, the more it contributes to improving online trust, digital governance, platform safety, and the integrity of public communication on social platforms, at least indirectly. A framework that can be implemented without collecting and sharing personally identifiable information is therefore especially pertinent to this objective, as it could enable platforms and researchers to work towards integrity without adding to the surveillance infrastructure of platforms working towards integrity. This dynamic has become increasingly apparent in discussions around content moderation and platform governance. It also extends to newly connected populations, since connectivity programmes aimed at rural and agricultural regions bring first-time users onto the same platforms without a corresponding expansion of moderation capacity (Yadav *et al.*, 2023).

Practical Implications

The architecture can support several other adjacent applications, beyond the sockpuppet detection use case that motivates it. Similarly, in the context of social media moderation, these four-modality patterns might also be used to flag coordinated inauthentic behavior beyond mere duplication, such as astroturfing campaigns or engagement farming. Temporal drift modeling might be used to detect sockpuppetry or credential sharing in online education platforms, which is similar to dormancy and bursts of coordinated activity. The same privacy-preserving design applies to digital identity management and public sector digital services, which are additional domains that are also subject to deduplication without the release of PII, like the context where Ahmed's (Ahmed, 2025) original architecture was applied, the CRM and healthcare contexts. At a more general level, late-fused, privacy-abstracted modalities could be used in other applications that demand privacy-preserving AI for behavioral integrity, such as pipelines for detecting misinformation that are required only to detect the presence of coordinated accounts without processing identifying content. Shared-access infrastructure domains raise closely related requirements, as privacy preservation alongside coordinated multi-tier access has been identified as a central open problem in shared-spectrum systems such as the Citizens Broadband Radio Service (Agarwal *et al.*, 2022). Deployment studies in agricultural and rural

connectivity settings report the same requirement arising where the user base is heterogeneous and the underlying infrastructure is shared (Yadav *et al.*, 2023).

These applications are offered as potential applications that fit the design of this framework; these are not necessarily proven applications, and each will need to be tested empirically before it can be adopted in practice.

Ethical Considerations

The framework is designed for real users, which means there are ethical issues beyond its technical aspects. A major concern is false positive: if an account is mistakenly identified as a sockpuppet, it could be suspended, shadow-moderated, or be subjected to reputational damage; if an account is not identified as a sockpuppet but the moderator believes it is, no harm is immediately done to the platform. This means that any deployment should involve an automated detection system coupled with manual review prior to any consequential action, especially in the initial, unvalidated use of the framework.

Architecturally, privacy is dealt with as outlined in Section VI.While architectural privacy preservation does reduce privacy risk, C is not the only such risk. In principle, behavior and time patterns, even when abstracted from the data, can be used with other data sources to re-identify individuals, and platforms that implement this framework are responsible for ensuring that the use of their outputs does not compromise the privacy protections built into the modalities.

Another factor to consider is algorithmic bias. The framework has not been evaluated for differential performance by the population and behavioral baselines (activity cadence, frequency of posting, dormancy patterns) are suggested to differ between cultural, linguistic, and access contexts and could result in differential false-positive rates when applied without population-specific calibration. Such access contexts are themselves shaped by infrastructure: connectivity initiatives in developing regions, including shared-spectrum deployments intended for rural and agricultural use (Yadav *et al.*, 2023), bring online user populations whose baseline activity patterns may differ markedly from those represented in existing datasets. Access technology itself varies across these contexts, as coverage is increasingly delivered through unlicensed and shared bands with different session, latency, and availability characteristics (Manekiya *et al.*, 2020; Agarwal *et al.*, 2022).

Lasttly, there must be transparency as to the theoretical status of the framework for responsible deployment. Presently, it has not been empirically benchmarked; thus, its presentation or systems developed from it as validated in production settings would overstate its current maturity. Empirical validation with preferably published performance metrics disaggregated by user population, and independent audit of false positive impact should precede any real-world use.

Limitations and Future Research

Because this is a theoretical contribution, there are no applications of it, there are no benchmarks, and there are no tests or experiences that have been validated. The architecture, comparisons, and constructs presented above are not yet empirical; the empirical work is described below. The most suitable initial evaluation benchmarks are the DARPA Twitter bot corpus and the Stanford Reddit sockpuppet dataset. Both corpora are drawn from mature, high-connectivity user populations, so calibration against populations reached through shared-spectrum and rural connectivity deployments remains an open requirement (Agarwal *et al.*, 2022; Yadav *et al.*, 2023). Currently, fusion weights are theory-motivated, but learning adaptive weights for fusion using supervised or semi-supervised learning would reinforce the model. The TDS L2 norm is the same for all behavioural dimensions, which should be further investigated in future work using feature-importance weighted drift vectors, as suggested by the dynamic concept-adaptation methods of Zhu *et al.* (2023). To enhance this architecture, cross-platform identity matching (same operator across multiple networks) and the fifth modality encoding social graph structure using graph neural networks (GNNs) are natural extensions. The stability of the device metadata modality under evolving access technologies is a further open question, given the continuing diversification of spectrum-sharing and unlicensed operation (Manekiya *et al.*, 2020). The use of explainable AI techniques, as shown to be feasible in neighbouring auditing situations (where adjacent auditing tasks share the same behavioural data but cannot be merged) would give practitioners the ability to audit individual detection decisions, not the fused output, and federated learning could facilitate cross-platform collaboration in the training of the model without the centralization of sensitive behavioral data.

CONCLUSION

This paper introduced a theoretical framework for privacy-preserving sockpuppet and duplicate profile detection in social networks based on multimodal approach, combining four different modalities. The formal extension of the late-fusion architecture with a Temporal Drift modality, which is defined by the Activity Window, Baseline Profile, Drift Vector, TDS, Dormancy Threshold and Burst Index, captures the dormancy-to-activation behavior that is structurally inaccessible with current static detection techniques. The framework is situated at the intersection of multimodal fusion, temporal sensitivity, privacy preservation, and unsupervised operation and has been validated through a theoretical comparison with previous works (Alharbi *et al.*, 2021; Li, 2024) that have analyzed and applied the problem of sybil detection in their own unique ways. It is grounded in previously established time-series anomaly theory and within the expanded literature on coordinated inauthentic behavior, concept drift, and trustworthy AI that have dealt with the same problem in their own unique ways (Section II). It gives a solid starting point for the future empirical research. In addition to its technical contribution, the framework is designed as an interdisciplinary basis that

links AI, social computing, and digital trust studies, and has practical relevance to platform safety, fighting against misinformation, and digital trust that respects privacy, until empirically validated and ethically protected as discussed in Sections IX and X.

REFERENCES

Agarwal, P., Manekiya, M., Ahmad, T., Yadav, A., Kumar, A., Donelli, M., & Mishra, S. T. (2022). A survey on Citizens Broadband Radio Service (CBRS). *Electronics, 11*(23), 3985. https://doi.org/10.3390/electronics11233985

Ahmed, M. O. S. (2025). A Late-Fusion Multimodal AI Framework for Privacy-Preserving Deduplication in National Healthcare Data Environments. 2025 *IEEE FMLDS*. https://doi.org/10.1109/FMLDS67896.2025.00021

Alharbi, A., Dong, H., Yi, X., Tari, Z., & Khalil, I. (2021). Social Media Identity Deception Detection. *ACM Computing Surveys, 54*(3), 1-35. https://doi.org/10.1145/3446372

Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., & Alomari, D. M. (2023). Machine learning-based social media bot detection: A comprehensive literature review. *Social Network Analysis and Mining, 13*(1). https://doi.org/10.1007/s13278-022-01020-3

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bellutta, D., & Carley, K. M. (2023). Investigating coordinated account creation using burst detection and network analysis. *J Big Data, 10*(1), 20. https://doi.org/10.1186/s40537-023-00695-7

Breuer, A., Khosravani, N., Tingley, M., & Cottel, B. (2023). Preemptive detection of fake accounts on social networks via multi-class preferential attachment classifiers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*.

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access, 11*, 78994-79015. https://doi.org/10.1109/access.2023.3294569

Cima, L., Mannocci, L., Avvenuti, M., Tesconi, M., & Cresci, S. (2024). Coordinated behavior in information operations on Twitter. *IEEE Access, 11*, 61568-61585.

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. *Decision Support Systems, 160*. https://doi.org/10.1016/j.dss.2022.113819

Cinus, F., Minici, M., Luceri, L., & Ferrara, E. (2024). Exposing Cross-Platform Coordinated Inauthentic *Activity in the Run-Up to the 2024 U.S. Election*. https://arxiv.org/abs/2410.22716

Ding, C., Zhao, J., & Sun, S. (2022). Concept Drift Adaptation for Time Series Anomaly Detection via Transformer. *Neural Processing Letters, 55*(3), 2081-2101. https://doi.org/10.1007/s11063-022-11015-0

Doshi, J., Novacic, I., Fletcher, C., Borges, M., Zhong, E., Marino, M. C., Gan, J., Mager, S., Sprague, D., & Xia, M. (2024). *Sleeper Social Bots: A New Generation of AI Disinformation Bots are Already a Political Threat*. https://arxiv.org/abs/2408.12603

Fu, D., He, J., Tong, H., & Maciejewski, R. (2022). *Privacy-Preserving Graph Analytics: Secure Generation and Federated Learning*. https://arxiv.org/abs/2207.00048

Goyal, B., Gill, N. S., Gulia, P., Prakash, O., Priyadarshini, I., Sharma, R., Obaid, A. J., & Yadav, K. (2023). Detection of Fake Accounts on Social Media Using Multimodal Data with Deep Learning. *IEEE Transactions on Computational Social Systems*.

Gruzd, A., Soares, F. B., & Mai, P. (2023). Trust and Safety on Social Media: Understanding the Impact of Anti-Social Behavior and Misinformation on Content Moderation and Platform Governance. *Social Media + Society, 9*(3). https://doi.org/10.1177/20563051231196878

Heeb, S., & Plesner, A. (2024). *Sybil Detection using Graph Neural Networks*. https://arxiv.org/abs/2409.08631

Jiang, H., Pei, J., Yu, D., Yu, J., Gong, B., & Cheng, X. (2023). Applications of Differential Privacy in Social Network Analysis: A Survey. *IEEE Transactions on Knowledge and Data Engineering, 35*(1), 108-127.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys, 55*(2), 1-38. https://doi.org/10.1145/3491209

Khaliq, A. R., Ullah, S., Ahmad, T., Yadav, A., & Majid, M. I. (2023). Behavioral analysis of backdoor malware exploiting heap overflow vulnerabilities using data mining and machine learning. *Pakistan Journal of Engineering, Technology & Science, 11*(1). https://doi.org/10.22555/pjets.v11i1.984

Khan, Z., Khan, Z., Lee, B.-G., Kim, H. K., & Jeon, M. (2024). Graph neural networks based framework to analyze social media platforms for malicious user detection. *Applied Soft Computing, 155*. https://doi.org/10.1016/j.asoc.2024.111416

Lazebnik, T., & Iny, O. (2024). Temporal graphs anomaly emergence detection: benchmarking for social media interactions. *Applied Intelligence, 54*(23), 12347-12356. https://doi.org/10.1007/s10489-024-05821-3

Li, J., Malialis, K., & Polycarpou, M. M. (2023). *Autoencoder-based Anomaly Detection in Streaming Data with Incremental Learning and Concept Drift Adaptation (strAEm++DD)*. https://arxiv.org/abs/2305.08977

Li, Y. (2024). *Malicious Social Bot Detection in Social Networks Based on Multi-modal Feature Fusion with Transformer Networks*. ACM CSCIT,

Liu, J., Yang, D., Zhang, K., Gao, H., & Li, J. (2023). Anomaly and change point detection for time series

with concept drift. *World Wide Web, 26*(5), 3229-3252. https://doi.org/10.1007/s11280-023-01181-z

Lu, H., Gong, D., Li, Z., Liu, F., & Liu, F. (2023). SybilHP: Sybil Detection in Directed Social Networks with Adaptive Homophily Prediction. *Applied Sciences, 13*(9). https://doi.org/10.3390/app13095341

Luceri, L., Salkar, T. V., & Balasubramanian, A. (2025). *Coordinated Inauthentic Behavior on TikTok: Challenges and Opportunities for Detection in a Video-First Ecosystem*. https://arxiv.org/abs/2505.10867

Manekiya, M., Kumar, A., Yadav, A., & Donelli, M. (2020). A survey of resource allocation techniques for cellular network's operation in the unlicensed band. *Electronics, 9*(9), 1464. https://doi.org/10.3390/electronics9091464

Mannocci, L. (2025). *Conceptual Framework and Methods for the Analysis of Coordinated Online Behavior National PhD Program in AI for Society*]. Pisa, Italy.

Mannocci, L., Mazza, M., Monreale, A., Vakali, A., & Tesconi, M. (2024). *Detection and Characterization of Coordinated Online Behavior: A Survey*. https://arxiv.org/abs/2408.01257

Moore, M. (2023). Fake accounts on social media, epistemic uncertainty and the need for an independent auditing of accounts. *Internet Policy Review, 12*(1).

Paganelli, M., Buono, F. D., Baraldi, A., & Guerra, F. (2022). Analyzing how BERT performs entity matching. *Proceedings of the VLDB Endowment, 15*(8), 1726-1738. https://doi.org/10.14778/3529337.3529356

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems, 263*. https://doi.org/10.1016/j.knosys.2023.110273

Saha, S., Ahmad, T., & Yadav, A. (2025). Miscellaneous applications of deep learning based multi-sensor Earth observation. In S. Saha (Ed.), *Deep learning for multi-sensor Earth observation* (Chapter 15). Elsevier.

Sunkara, S. P. (2026a). AI-Enabled Drone Image Analysis Workflows for Condition Monitoring of Power Distribution Assets. *2026 International Symposium of Systems, Advanced Technologies and Knowledge (ISSATK), Hammamet, Tunisia,* 1–6. https://doi.org/10.1109/ISSATK69667.2026.11566851

Sunkara, S. P. (2026b). Enhancing Grid Stability with AI-Driven Control Systems. *2026 International Symposium of Systems, Advanced Technologies and Knowledge (ISSATK), Hammamet, Tunisia.* https://doi.org/10.1109/ISSATK69667.2026.11566761

Sunkara, S. P. (2026c). Physics-Informed AI Digital Twins For Ultra-Fast, Scalable Time-Series Power Flow And Automated Grid Constraint Resolution. *International Journal of Advances in Signal and Image Sciences, 12*(3s), 356–377. https://doi.org/10.29284/ecrg7s57

Sunkara, S. P., & Sunkara, K. (2026). Digital Twin–Enabled Grid Resilience: Predictive Failure Analysis and Proactive Control for Climate-Driven Power System. *International Journal of Computer Information Systems and Industrial Management Applications, 18*(1s), 479–491. https://doi.org/10.70917/ijcisim-2026-1940

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447-464. https://doi.org/10.1007/s12525-020-00441-4

Tsikerdekis, M., & Zeadally, S. (2014). Multiple Account Identity Deception Detection in Social Media Using Nonverbal Behavior. *IEEE Transactions on Information Forensics and Security, 9*(8), 1311-1321. https://doi.org/10.1109/TIFS.2014.2325052

Xiang, Q., Zi, L., Cong, X., & Wang, Y. (2023). Concept Drift Adaptation Methods under the Deep Learning Framework: A Literature Review. *Applied Sciences, 13*(11). https://doi.org/10.3390/app13116515

Yadav, A., Manekiya, M., Ahmad, T., & Khan, M. A. (2023). Adopting sustainable agriculture practices in developing nations using CBRS. *2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences* (ICETAS).

Yang, W., Wei, Y., & Wei, H. (2023). Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems, 3*, 161-188.

Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., & Salehi, M. (2024). Deep Learning for Time Series Anomaly Detection: A Survey. *ACM Computing Surveys, 57*(1), 1-42. https://doi.org/10.1145/3691338

Zhong, C., & Goel, S. (2024). Transparent AI in Auditing through Explainable AI. *Current Issues in Auditing, 18*(2), A1-A14. https://doi.org/10.2308/ciia-2023-009

Zhu, J., Cai, S., Deng, F., Ooi, B. C., & Zhang, W. (2023). *METER: A Dynamic Concept Adaptation Framework for Online Anomaly Detection*. https://arxiv.org/abs/2312.16831

Zouzou, Y., & Varol, O. (2024). Unsupervised detection of coordinated fake-follower campaigns on social media. *EPJ Data Science, 13*.

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CIB	Coordinated Inauthentic Behavior
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GDPR	General Data Protection Regulation
GNN	Graph Neural Network
HIPAA	Health Insurance Portability and Accountability Act
PCA	Principal Component Analysis
PII	Personally Identifiable Information
TDS	Temporal Drift Score
XAI	Explainable Artificial Intelligence