Available on: http://gjetr.ep-journals.org/ E-ISSN: 3051-3782, P-ISSN: 3051-3774

Global Journal of Engineering and Technology Review

Volume: 02 Issue: 09 September 2026

Semantic De-Duplication of Shared Cyber Threat Indicators using Large Language Model Embeddings

Abolaji Adebayo
Computacenter, USA

Chukwunenye Amadi
Independent Researcher, USA

Ayokunle Olamide Ijagbemi
Independent Researcher, USA

Published Date: September 05, 2026 **Article DOI : 10.65150/EP-gjetr/V2E9/2026-07**

ABSTRACT: As defence agencies, allied partners, and cybersecurity firms exchange threat intelligence, the same indicator is frequently reported many times in slightly altered form, producing large-scale duplication across intelligence repositories. Traditional de-duplication relies on exact matching using hashes or literal text comparison and fails when information is reworded, paraphrased, or contextually recast. This paper proposes a semantic de-duplication approach that uses Large Language Model (LLM) embeddings to identify duplicate threat indicators on the basis of meaning rather than form. The method encodes each report into a contextual vector, measures pairwise similarity, clusters semantically equivalent reports, and retains a single representative entry per cluster. Situated within the literature on record linkage, entity resolution, and near-duplicate detection, the paper specifies an evaluation protocol based on precision, recall, F1-score, cluster-level agreement, and latency over open-source and synthetic corpora, together with the baselines and reproducibility requirements against which the design should be tested. That protocol is defined here but has not yet been executed: no dataset has been processed and no numerical results are reported. The behaviour described in the discussion is the behaviour the design is expected to exhibit, offered as hypotheses to be tested rather than as measured findings, and the anticipated benefits of consolidation, namely reduced noise, less exaggerated risk assessment, and lower computational and storage cost, are stated in those terms throughout. The paper contributes a meaning-based de-duplication design for defence CTI, an evaluation protocol for validating it, and a discussion of the accuracy, efficiency, and governance implications of deploying it.

KEYWORDS: semantic de-duplication, near-duplicate detection, record linkage, embeddings, cosine similarity, cyber threat intelligence, resource optimization.

I. INTRODUCTION

The routine exchange of cyber threat intelligence (CTI) has become a foundational practice in contemporary network defence, transforming isolated observations of malicious activity into a shared resource that many organizations can act upon collectively (Dosunmu & Ogundele, 2026a; Jimoh & Ahmed, 2024; Ogunwola, 2026). Indicators of compromise, including file hashes, domain names, network addresses, uniform resource locators, and the descriptive context that accompanies them, circulate through community platforms, commercial feeds, and governmental advisories at a volume that no single analyst could manually reconcile (Ogunwola & Alozie, 2026a; Ogunwola et al., 2026). As participation in these ecosystems has widened, the raw quantity of shared artifacts has grown far faster than the human capacity to curate it (Ogunwola & Deborah, 2026). The promise of collective defence rests on the assumption that pooling observations yields a clearer and more timely picture of adversary behaviour than any participant could assemble in isolation. That assumption, however, is complicated by an unglamorous but pervasive problem: the same underlying piece of intelligence is frequently reported many times, in many forms, by many contributors. What appears to be a rich and diverse corpus is often a far smaller set of distinct facts wrapped in redundant packaging. This paper concerns the identification and consolidation of such redundancy at the semantic level, where two records may describe an identical threat while sharing few or none of the exact tokens that conventional matching relies upon. Addressing this gap is a precondition for the scalable, trustworthy intelligence sharing that defensive communities increasingly depend upon (Adegbite et al., 2020, 2023).

Duplication in shared intelligence is not merely an aesthetic inconvenience; it degrades the operational value of the very feeds intended to strengthen defence. When a single campaign is described under several slightly different phrasings, downstream analytics inflate its apparent prevalence, distort prioritization, and consume scarce analyst attention with repeated triage of what is effectively one event. Automated enrichment pipelines and correlation engines amplify these distortions, propagating redundant entries into tickets, dashboards, and blocking rules. The consequence is a paradox in which the abundance meant to inform defenders instead overwhelms them, and confidence in the underlying feed erodes as consumers encounter the same intelligence repeatedly under different guises. The difficulty is compounded by the textual and

heterogeneous nature of much shared context: free-form descriptions, analyst commentary, and narrative reports that resist the rigid schemas assumed by traditional deduplication. Two contributors observing the same phishing infrastructure may record entirely different sentences, cite different observable subsets, and adopt different naming conventions for the same adversary, yet convey the same actionable meaning. Recognizing that these records are semantically equivalent demands a notion of similarity grounded in meaning rather than surface form. It is precisely this class of near-duplicate that has proven most resistant to established tooling, and that motivates the representation-centric approach developed in the remainder of this work.

The central premise of this paper is that recent advances in dense semantic representation, particularly embeddings derived from large language models, offer a principled foundation for detecting duplication that eludes exact and lexical methods (Adebayo, Adegbite, & Ahmed, 2023; Hussain & Adebayo, 2026). By projecting heterogeneous threat descriptions into a continuous vector space in which proximity reflects meaning, such representations make it possible to recognize that two dissimilarly worded records refer to the same underlying phenomenon. This reframes deduplication as a problem of measuring semantic distance and clustering near neighbours, rather than of matching literal strings or normalizing rigid fields. The approach inherits both the strengths and the risks of learned representations: it captures paraphrase, synonymy, and structural variation that defeat token-level comparison, but it also introduces sensitivity to domain shift, the possibility of conflating genuinely distinct indicators that merely sound alike, and non-trivial computational cost at scale. A responsible treatment must therefore situate embedding-based matching within the longer intellectual tradition of record linkage and near-duplicate detection, drawing on decades of methodological insight about blocking, thresholding, and evaluation. Only by connecting the new representational machinery to these established foundations can one design a deduplication pipeline that is both semantically expressive and operationally dependable, and that degrades gracefully when confronted with the ambiguity endemic to real-world intelligence sharing rather than failing silently or conflating unrelated threats.

This paper makes several contributions organized across its two halves. It first characterizes the phenomenon of duplication in shared threat intelligence, articulating its causes and its operational consequences, and demonstrating why exact and lexical matching are structurally inadequate for the semantic near-duplicates that dominate practical corpora. It then consolidates the conceptual foundations on which a semantic solution must rest, synthesizing the literatures of record linkage and entity resolution, of string similarity and near-duplicate detection, and of neural representation learning and scalable similarity search into a coherent basis for the method that follows. Finally, it positions the problem within the broader landscape of applied cyber defence and cross-sector intelligent systems, clarifying how deduplication relates to the wider goals of threat intelligence sharing, mining, and cybersecurity data science. Throughout, the emphasis is on meaning-preserving consolidation: reducing redundancy without discarding the corroborating signal that multiple independent reports of the same threat legitimately provide. The remaining sections develop these themes in turn, beginning with a detailed account of how duplication arises in sharing ecosystems and why it resists conventional treatment, before turning to the representational and algorithmic machinery that enables a semantic response. The result is intended both as a practical design and as a bridge between mature data-matching theory and the emerging capabilities of language-model embeddings.

II. DUPLICATION IN SHARED THREAT INTELLIGENCE

A. Causes and Consequences of Duplication

Duplication in shared threat intelligence arises from the structure of the sharing ecosystem itself rather than from any single point of failure. A given malicious campaign is typically observed by many organizations, each of which may report its findings independently to overlapping communities and feeds. The same domain, sample, or infrastructure is thus described repeatedly, often with subtle variation in the accompanying context, the subset of observables included, and the vocabulary used to characterize the adversary. Aggregation platforms compound the effect by ingesting from numerous upstream sources, some of which themselves re-share intelligence obtained elsewhere, producing cascades in which a single original observation reappears many times as it propagates (Agu et al., 2023b; Akinleye & Adeyoyin, 2021). Temporal dynamics add a further layer: as an incident unfolds, successive reports update, correct, and extend earlier ones, so that a chronological series of near-identical records accumulates around one evolving event. Differences in tooling and schema mean that identical facts are serialized differently, while free-text commentary introduces essentially unbounded surface variation. The absence of universal identifiers for adversaries, campaigns, and infrastructure removes the natural anchor that would otherwise permit trivial consolidation (Orise & Niniola, 2025). The net result is a corpus in which apparent volume substantially overstates the count of distinct underlying facts, and in which the redundancy is distributed unevenly, concentrated around high-profile threats that attract the most independent reporting and therefore the most duplication.

The consequences of unmanaged duplication extend across the analytic and operational lifecycle. At the level of situational awareness, repeated reports of a single threat inflate prevalence estimates and skew prioritization, causing defenders to over-weight whatever is most frequently re-shared rather than what is most dangerous or novel. Analyst productivity suffers directly, as scarce expert attention is expended re-triaging intelligence that has already been assessed, and as the cognitive burden of distinguishing genuine corroboration from mere echo grows with corpus size. Automated pipelines translate these distortions into concrete costs: redundant entries expand storage and indexing requirements, multiply enrichment lookups against rate-limited external services, and generate duplicate alerts, tickets, and blocking rules that clutter downstream systems (Anene & Clement, 2022, 2024). Perhaps most corrosively, pervasive duplication erodes trust in the feed itself, as consumers who repeatedly encounter the same intelligence under different guises come to discount its reliability and, in the worst case, disengage from sharing altogether. Yet duplication cannot simply be eliminated wholesale, because multiple independent reports of the same threat carry legitimate corroborative value, raising confidence in an observation and enriching it with complementary context. The objective, therefore, is not indiscriminate removal but meaning-preserving consolidation that collapses redundant descriptions into a single enriched record while retaining the evidentiary weight that independent confirmation provides, a balance that any effective deduplication method must be designed explicitly to strike.

B. Limits of Exact and Lexical Matching

The most immediate response to duplication is exact matching, in which records are declared identical only when their canonical representations coincide byte for byte. This strategy is fast, deterministic, and appropriate for atomic observables such as cryptographic hashes, where any difference genuinely signals a different artifact. It fails, however, the moment meaning is carried by anything other than a rigid identifier. Trivial

normalization differences, such as case, whitespace, punctuation, ordering of fields, or the presence of defanged notations, defeat literal comparison, and free-text context defeats it almost entirely, since two analysts virtually never phrase the same observation identically. Lexical similarity methods relax this rigidity by scoring partial overlap. Edit distance quantifies the number of character operations separating two strings (Levenshtein, 1966), and broader families of string distance metrics extend this to token-level and phonetic comparison (Cohen et al., 2003), tolerating typographic drift and minor reformatting. For efficiency at scale, near-duplicate detection has long relied on hashing schemes that estimate set resemblance and containment through sampled sketches (Broder, 1997) or that map similar inputs to nearby binary codes so that collisions approximate similarity (Charikar, 2002). These techniques underpin large-scale document deduplication and remain valuable where redundancy manifests as surface overlap, offering sub-linear candidate generation without exhaustive pairwise comparison across the corpus.

The limitation shared by all of these methods is that they measure form rather than meaning. Sketch-based resemblance (Broder, 1997) and similarity-preserving hashing (Charikar, 2002) presuppose substantial overlap in tokens or shingles; when two records describe the same threat using disjoint vocabulary, whether through different sentence structures, synonymous terminology, distinct observable subsets, or divergent naming for the same adversary, the shared surface shrinks toward zero and the estimated similarity collapses, even though the underlying meaning is identical. Edit distance (Levenshtein, 1966) behaves analogously, reporting large character-level divergence for paraphrases that a human would immediately recognize as equivalent, while string metric ensembles (Cohen et al., 2003) mitigate but do not escape this dependence on lexical proximity. The failure is not one of tuning but of representation: methods that operate on characters and tokens are blind to semantics by construction, so no threshold setting can make them recognize meaning-preserving variation without simultaneously admitting spurious matches between superficially similar but genuinely distinct records. Conversely, tightening thresholds to suppress false matches suppresses precisely the paraphrastic near-duplicates that dominate real intelligence corpora. This structural inadequacy motivates a shift from surface comparison to representations in which proximity reflects meaning. Rather than discarding lexical techniques, an effective pipeline retains them for what they do well, namely exact and near-verbatim matching and cheap candidate blocking, while delegating the semantic judgement they cannot make to learned representations, which the following section grounds in the broader theory of matching and representation.

III. FOUNDATIONS: RECORD LINKAGE, SIMILARITY, AND REPRESENTATION

A. Record Linkage and Entity Resolution

The problem of deciding whether distinct records refer to the same real-world entity long predates threat intelligence, and its methodological literature supplies the scaffolding on which semantic deduplication is built. Duplicate record detection has been studied extensively as a canonical data-quality task, with surveys cataloguing the techniques used to identify and resolve records that describe the same entity despite representational differences (Elmagarmid et al., 2007). The broader discipline of data matching and record linkage formalizes this as a pipeline of preprocessing, blocking, comparison, and classification, in which pairs of records are compared across multiple fields and assigned a match status on the basis of aggregated similarity (Christen, 2012). Blocking, in particular, is a recurring concern: because exhaustive pairwise comparison scales quadratically, the field has developed indexing and partitioning strategies that restrict comparison to plausible candidates, a principle that carries directly into embedding-based systems where approximate nearest-neighbour search plays the analogous role. Comparative evaluations of entity-matching frameworks have examined how alternative combinations of blocking, similarity functions, and decision models perform across datasets, exposing the sensitivity of results to configuration and the difficulty of transferring a tuned system to a new domain (Köpcke & Rahm, 2010). These insights transfer intact to the intelligence setting, where the entities are threats and campaigns rather than customers or citations, but the underlying resolution problem is structurally the same and the accumulated engineering wisdom remains applicable.

Beyond the deterministic and rule-based traditions, probabilistic and statistical formulations provide a principled account of uncertainty in matching decisions. Bayesian approaches to record linkage and de-duplication treat the assignment of records to latent entities as an inference problem, jointly estimating which records coreference and propagating the resulting uncertainty into downstream estimates rather than committing prematurely to hard clusters (Steorts et al., 2016). This perspective is valuable for threat intelligence precisely because semantic equivalence is often genuinely ambiguous: two reports may plausibly describe the same campaign or two related but distinct ones, and a system that quantifies rather than conceals this ambiguity is better suited to high-stakes defensive use. The record-linkage tradition also emphasizes evaluation discipline, insisting that matching quality be assessed against curated ground truth with attention to both precision and recall, and warning against the class imbalance that makes naive accuracy misleading when true duplicates are rare relative to all candidate pairs (Christen, 2012). It further foregrounds the transitivity problem, whereby pairwise match decisions must be reconciled into globally consistent entity clusters, a step that embedding-based clustering inherits directly. Taken together, these foundations establish that deduplication is not a single similarity computation but an end-to-end process of candidate generation, comparison, decision, and consolidation, each stage of which the semantic method must instantiate (Elmagarmid et al., 2007; Köpcke & Rahm, 2010).

B. String Similarity and Near-Duplicate Detection

Situated between the record-linkage tradition and modern representation learning lies a substantial body of work on measuring textual similarity, which supplies both the vocabulary and the baseline methods against which semantic approaches are judged. Surveys of text similarity organize the field into character-based, term-based, corpus-based, and knowledge-based families, clarifying that the notion of similarity is not monolithic but depends on whether one compares surface strings, weighted term distributions, or learned semantic representations (Gomaa & Fahmy, 2013). The term-based family is anchored in the vector space model, in which documents are represented as sparse vectors over a vocabulary and similarity is computed as the cosine of the angle between them. Central to this model is the weighting of terms so that discriminative words contribute more than ubiquitous ones, an insight formalized in the family of term-frequency and inverse-document-frequency schemes that remain a durable baseline for textual comparison (Salton & Buckley, 1988). These representations are efficient, interpretable, and effective when shared meaning coincides with shared vocabulary, and they underpin much of classical information retrieval, where ranking documents by similarity to a query is the defining operation (Manning et al., 2008). For threat descriptions with substantial lexical overlap, such weighted term models already improve markedly on naive matching.

The enduring weakness of sparse term-based similarity is the vocabulary mismatch problem: because each term occupies an independent dimension, the model treats synonyms as wholly unrelated and cannot recognize that differently worded descriptions convey the same meaning (Manning et al., 2008). Weighting refinements (Salton & Buckley, 1988) and the various string and set-based measures catalogued in the similarity literature (Gomaa & Fahmy, 2013) mitigate but do not resolve this, because they remain grounded in the observed tokens rather than in the concepts those tokens denote. Information retrieval has historically addressed the gap through techniques such as query expansion, relevance feedback, and latent decomposition of the term-document matrix, each attempting to bridge surface forms that share meaning (Manning et al., 2008). These partial remedies foreshadow the central move of dense representation learning, which is to replace the sparse, orthogonal term space with a lower-dimensional continuous space in which semantically related expressions are placed near one another by construction. In the deduplication setting, the practical implication is a division of labour: sparse lexical similarity remains an inexpensive and reliable signal for near-verbatim redundancy and for generating candidate pairs cheaply, while the residual and most valuable cases, those records that are semantically identical yet lexically disjoint, require the meaning-aware representations developed next. The similarity literature thus frames dense embeddings not as a replacement for classical measures but as the component that supplies the semantic judgement those measures structurally lack, completing rather than superseding the toolkit.

C. Embeddings and Scalable Similarity Search

Dense distributed representations address vocabulary mismatch by embedding textual units into continuous vector spaces in which geometric proximity encodes semantic relatedness. Early neural word embeddings learned such spaces from co-occurrence statistics, producing vectors whose arithmetic captured meaningful linguistic regularities (Mikolov et al., 2013), while related global factorization of co-occurrence counts yielded comparably expressive word vectors (Pennington et al., 2014). These static embeddings, however, assign a single vector to each word irrespective of context, limiting their fidelity for polysemous terms and full sentences. The introduction of the transformer architecture, built on self-attention, enabled contextual representations in which a token's vector depends on the surrounding text (Vaswani et al., 2017), and pretrained bidirectional encoders learned deep contextual embeddings that substantially advanced language understanding (Devlin et al., 2019). The same architectural foundation underlies large generative language models whose scale confers broad few-shot capability and rich internal representations (Brown et al., 2020). For similarity tasks specifically, sentence-level encoders were developed to produce fixed-length embeddings whose cosine proximity reflects semantic equivalence, making pairwise comparison efficient (Reimers & Gurevych, 2019); universal sentence encoders pursued the same goal of general-purpose sentence vectors (Cer et al., 2018); and contrastive objectives further sharpened the isotropy and discriminative quality of sentence embeddings (Gao et al., 2021). Collectively these representations supply exactly the meaning-aware geometry that lexical methods lack, mapping paraphrastic threat descriptions to nearby points regardless of their surface divergence.

Possessing a semantic geometry is necessary but not sufficient; deduplication over large corpora further requires that near neighbours be found and grouped efficiently, since exhaustive pairwise comparison scales quadratically and is infeasible at operational volume. Approximate nearest-neighbour search provides the blocking analogue for dense spaces: graph-based indexes organize vectors into navigable small-world structures that answer proximity queries in logarithmic time with high recall (Malkov & Yashunin, 2020), and library-scale systems combine such indexes with quantization to support billion-scale similarity search on commodity hardware (Johnson et al., 2021). Retrieved neighbours must then be consolidated into entity clusters, and density-based clustering is well suited to this because it discovers clusters of arbitrary shape and explicitly labels isolated points as noise rather than forcing every record into a group (Ester et al., 1996); hierarchical density-based extensions relax the need to fix a single global density threshold, accommodating the uneven concentration of duplication across a corpus (McInnes et al., 2017). Broader surveys of text clustering situate these choices within the space of algorithms and representations available for grouping documents by content (Aggarwal & Zhai, 2012), while generative topic models offer a complementary probabilistic view of thematic structure that can inform corpus organization (Blei et al., 2003). Together, semantic embeddings, approximate nearest-neighbour indexing, and density-based clustering compose the end-to-end machinery of representation, candidate generation, and consolidation that the record-linkage tradition prescribes, now instantiated for meaning rather than surface form.

IV. RELATED WORK: APPLIED CYBER DEFENCE AND CROSS-SECTOR INTELLIGENT SYSTEMS

Relationship to the Authors' Prior Work

This manuscript is a companion to earlier work by the present authors on the same application area, and the relationship between the two is stated here explicitly so that readers and reviewers can assess what is new. That earlier study addresses real-time semantic correlation and de-duplication of shared threat indicators using large language models and natural language processing (Adebayo, Amadi, & Ijagbemi, 2026). Its scope is architectural: it advances a conceptual framework in which correlation across sources, extraction of structure from narrative reporting, and consolidation of redundant entries appear as stages of a single real-time pipeline, and its principal concerns are the latency, throughput, and governance constraints such a pipeline must satisfy. De-duplication appears there as one component among several, characterized at the level of system architecture rather than of method.

The present manuscript narrows to that component and treats it as an object of study in its own right, and it differs from the earlier work along four axes. In originality and framing, it recasts de-duplication not as a pipeline stage but as a problem in meaning-level record linkage, and grounds it in the record-linkage, entity-resolution, and near-duplicate-detection literatures surveyed in Section III, a foundation the earlier work does not develop. In methodology, it commits to a specific four-stage design, comprising contextual encoding, thresholded cosine similarity over an approximate nearest-neighbour index, density-based clustering, and reversible representative selection with an audit trail, and it specifies the merge semantics, the threshold-selection procedure, and the provenance-preserving record format that the earlier architectural treatment leaves open. In evaluation, it defines a protocol of its own, set out in Section V.E, built around pairwise detection metrics, cluster-level agreement, redundancy reduction, and latency, together with the lexical and syntactic baselines a semantic method must be measured against.

In datasets and experiments there is no overlap to disclose, because neither manuscript reports a completed empirical evaluation: no corpus, annotation, result, figure, or table is carried over from the earlier study to this one, and the evaluation protocol described in Section V.E is specific

to the de-duplication problem addressed here. The contribution claimed for the present work is therefore a self-contained design for meaning-based consolidation, a protocol for validating it, and an analysis of its accuracy, efficiency, and governance implications, rather than an extension of results previously reported. Textual overlap between the two manuscripts is confined to the framing of the threat-intelligence sharing problem that both share, and the earlier work is cited wherever that framing is drawn upon.

The methodological foundations above acquire their significance within the applied context of cyber threat intelligence sharing, a domain whose infrastructure and practices define both the need for deduplication and the constraints under which it must operate. Community sharing platforms have become the connective tissue of collective defence, enabling structured exchange of indicators and context among trust groups (Wagner et al., 2016), and surveys of technical threat intelligence map the sources, formats, and processing pipelines through which such intelligence flows (Tounsi & Rais, 2018). Studies of sharing practice examine the incentives, standards, and organizational dynamics that govern participation and quality (Wagner et al., 2019), while work on collective cyber defence articulates how pooled observation can strengthen the security posture of an entire community when redundancy and noise are managed rather than amplified (Skopik et al., 2016). Empirical examination of sharing platforms has documented their proliferation, heterogeneity, and the practical data-quality challenges, including duplication, that limit their effectiveness (Sauerwein et al., 2017). In parallel, the maturing field of CTI mining surveys the automated extraction and analysis of intelligence from diverse sources (Sun et al., 2023), and the broader programme of cybersecurity data science frames the application of machine learning and analytics to defensive problems, of which semantic consolidation of shared indicators is a natural instance (Sarker et al., 2020). Positioned against this landscape, the present work applies the intelligent-systems methods common across data-driven sectors to the specific, consequential problem of meaning-level redundancy in threat intelligence.

The framework advanced in this paper is positioned against a broad and rapidly growing body of applied scholarship. Three strands are especially pertinent: research on cyber defence and threat intelligence, which defines the operational problem; research on artificial intelligence and secure digital systems, which supplies the analytical machinery; and a wider literature that applies data-driven modelling, risk governance, and intelligent decision support across other critical-infrastructure and enterprise sectors, which demonstrates the breadth and transferability of the methods on which real-time semantic processing draws. The remainder of this section surveys these strands in turn.

A mature stream of applied research addresses the detection, prioritization, and containment of evolving threats in enterprise and national settings. This work develops maturity models, threat-actor analysis, incident-response and security-orchestration frameworks, and threat-informed defence engineering that together define the operational requirements the present study seeks to accelerate: rapid triage of high-volume indicators, measurable control effectiveness, and coordinated response across organizational boundaries (Adebayo, Adepoju, & Ozowara, 2023; Adegbite et al., 2024a, 2025; Ahmed et al., 2021; Dosunmu & Ogundele, 2021, 2022, 2023, 2024a, 2024b, 2024d, 2025a, 2025c).

Complementary studies examine the analytical capabilities that artificial intelligence and machine learning bring to security and decision support, including predictive analytics, human-in-the-loop annotation, generative models, edge intelligence, and interpretable learning. Collectively they establish both the promise of automated inference over large, noisy datasets and the necessity of retaining human oversight where consequences are severe (Adelanwa et al., 2023a; Adeyelu, 2020; Adeyelu & Dagodzo, 2024a; Afrihyia et al., 2025; Akanbi et al., 2025; Aliliele et al., 2023c; Annan, 2024; Atakpa & Abetoh, 2022; Borah et al., 2022; Dagodzo et al., 2022; Eboh & Aliliele, 2025b; Fawehinmi et al., 2026).

A substantial governance literature specifies the control, oversight, and accountability conditions under which intelligent systems must operate, spanning regulatory compliance, data-protection and privacy engineering, identity and access management, continuous auditing, and audit-automation. These contributions are directly relevant to fielding language models in defence, where transparency, provenance, and role-based control are prerequisites rather than afterthoughts (Adebayo et al., 2025; Adeyelu, 2018, 2021, 2022; Adeyelu & Dagodzo, 2022).

Work on cloud engineering, secure software delivery, and enterprise integration informs the engineering foundations required to deploy language-based analytics reliably at scale, addressing configuration governance, integration patterns, secure testing, and the operational discipline needed to keep data pipelines dependable under load (Aliliele et al., 2023a; Aminu-Ibrahim et al., 2019; Badmus et al., 2018).

Beyond the security domain, research on supply-chain optimization, procurement, and logistics demonstrates closely analogous data-driven approaches to coordinating, de-duplicating, and governing information across distributed, multi-stakeholder operations. The problems of reconciling heterogeneous records, eliminating redundant entries, and maintaining a single authoritative view mirror those confronted in threat-intelligence consolidation (Agbabiaka et al., 2019; Ahiaeke Patrick et al., 2020, 2021). A closely related body of work on pharmaceutical distribution, access, and quality assurance examines how records of products, prescriptions, and distribution events are reconciled across fragmented custody chains, where duplicate or inconsistent entries bear directly on availability and patient safety, and where systematic screening and quality frameworks are used to establish when two records describe the same underlying item (Asiedu, 2022, 2023, 2024, 2025, 2026; Asiedu & Asiedu, 2022, 2023a, 2023b, 2024a, 2024b).

Parallel work on energy systems, smart grids, and renewable infrastructure applies predictive modelling, continuous monitoring, and resilience engineering to complex critical-infrastructure environments analogous to those defended here, offering transferable insights on real-time inference, degradation detection, and the equitable, dependable operation of systems under stress (Dagodzo, 2018a; Komi, 2024). Work on energy-harvesting and intermittently powered wireless systems approaches the same resource-bounded inference problem from the hardware side, characterizing energy, reliability, latency, and throughput trade-offs under strict operating budgets and applying predictive power management to sustain dependable service, considerations that mirror the cost of encoding and indexing every incoming indicator in a continuously ingesting pipeline (Asiedu & Quainoo, 2023a, 2023b, 2024, 2025; Asiedu et al., 2025, 2026; Oyeleke et al., 2026; Quainoo et al., 2024, 2025a, 2025b).

Studies of industrial and occupational safety contribute mature models of hazard detection, near-miss signal analysis, and operational risk governance. Their central concern, distinguishing weak leading signals from background noise before an incident occurs, directly parallels the signal-versus-noise framing that motivates semantic filtering and prioritization in this study (Adeyelu, 2019, 2026). Research on non-destructive testing, weld integrity, and inspection standards is instructive for a further reason: it formalizes how an artifact can be examined and validated without altering or damaging it, together with the calibration and standardization such examination demands, which is precisely the property that

reversible, audit-preserving consolidation seeks in its treatment of shared indicators (Atta, 2026a, 2026b; Atta et al., 2018; Atta et al., 2020; Atta et al., 2021; Atta et al., 2022; Atta et al., 2024; Atta et al., 2025).

Research on healthcare and diagnostic infrastructure offers further evidence on lifecycle management, resilience planning, and performance measurement for mission-critical systems that must remain dependable under public-health stress, reinforcing the importance of governance, standardization, and measurable outcomes when scaling sensitive systems (Aminu-Ibrahim & Ogbete, 2023).

Contributions on geospatial intelligence, UAV and remote-sensing systems, and aviation-safety oversight demonstrate how heterogeneous sensor and spatial data are fused into actionable operational intelligence and how regulatory oversight is exercised over autonomous platforms, a fusion-and-oversight challenge closely related to correlating multi-source threat indicators (Adegbite et al., 2024b; Adeyelu, 2023).

A body of work on financial analytics, fraud detection, and business-process optimization shows how analytical models detect anomalies, reduce redundancy, and streamline decision workflows in high-stakes, data-intensive settings, providing methodological parallels for the anomaly-sensitive, redundancy-averse processing pursued in defence intelligence (Edivri & Oteri, 2022).

Further contributions on data analytics, business intelligence, and continuous monitoring illustrate how meaning and actionable structure are extracted from large, heterogeneous data streams for timely decision making (Akin-Oluyomi et al., 2020).

V. A METHOD FOR SEMANTIC DE-DUPLICATION

A. Method Overview

The proposed method treats de-duplication of shared cyber threat indicators as a pipeline that transforms free-form, heterogeneously worded observations into a compact set of canonical entries, each standing for a cluster of semantically equivalent reports. The pipeline proceeds in four coordinated stages that are deliberately kept modular so that operators may substitute components without disturbing the surrounding logic. In the first stage, each incoming indicator, together with its surrounding descriptive context, is normalized and encoded into a dense vector that captures meaning rather than surface form. In the second stage, these vectors are compared under a similarity measure, and a tunable threshold together with a clustering procedure partitions the corpus into groups of near-duplicates. In the third stage, a single representative is selected from each group and the remaining members are merged into it, with their provenance, confidence, and temporal metadata preserved rather than discarded. In the fourth stage, the consolidated records are emitted alongside an audit trail that allows any merge decision to be reviewed and, if necessary, reversed. This separation of concerns permits the encoding model, the index, and the clustering algorithm to evolve independently as the threat landscape and the underlying tooling mature. Table I summarizes the technical components associated with each stage and the principal design choices that govern their behaviour in an operational deployment.

Table I. Data-management techniques applied to semantic de-duplication

Technique	Description	Purpose
Contextual embeddings	Encode meaning as vectors	Enable semantic comparison
Cosine similarity	Measure vector closeness	Identify duplicates
Approximate nearest neighbour	Scalable similarity search	Support real-time operation
Clustering	Group equivalent reports	Collapse duplicates
Precision, recall, F1	Accuracy measures	Validate de-duplication quality

B. Semantic Encoding

The encoding stage is responsible for mapping each indicator and its accompanying narrative into a fixed-length vector whose geometry reflects semantic proximity. Contextualized language models built on the attention mechanism (Vaswani et al., 2017) provide the foundation for this mapping, because they resolve the meaning of a token from the words that surround it rather than from a static lookup, allowing paraphrases, abbreviations, and reordered descriptions to converge toward similar representations. Masked language modelling (Devlin et al., 2019) supplies pretrained parameters that already encode a broad understanding of technical and general English, which is valuable when threat reports mix jargon, vendor nomenclature, and informal prose. However, the raw output of such models is poorly suited to direct similarity comparison, since the vectors it produces are not calibrated for distance-based matching. Siamese and triplet fine-tuning (Reimers & Gurevych, 2019) addresses this by shaping an embedding space in which cosine distance is a faithful proxy for semantic relatedness, and contrastive self-supervision (Gao et al., 2021) further sharpens that space by pulling paraphrases together and pushing unrelated statements apart without requiring extensive manual labels. Where lightweight deployment or multilingual coverage is a priority, a general-purpose sentence encoder (Cer et al., 2018) offers a competitive alternative with modest computational demands. The method remains agnostic to the specific encoder, requiring only that the chosen model expose a stable, normalized vector suitable for downstream indexing.

C. Similarity, Thresholding, and Clustering

Once indicators are embedded, the method must decide which pairs are close enough to be treated as duplicates. Cosine similarity between normalized vectors serves as the primary measure, chosen because it is invariant to vector magnitude and because the encoders described above are trained to make angular proximity meaningful. A naive approach would compare every vector against every other, but the quadratic cost of exhaustive comparison is prohibitive for feeds that accumulate millions of indicators. To make the search tractable, the method delegates neighbour retrieval to an approximate nearest-neighbour index (Johnson et al., 2021), which returns the most similar candidates for each query in sublinear time while trading a controlled and typically negligible amount of recall for large gains in throughput. Graph-based indexing (Malkov & Yashunin, 2020) offers a complementary structure whose navigable small-world organization sustains low query latency even as the corpus grows and its intrinsic dimensionality rises. The retrieval step therefore narrows the field to a short list of plausible matches per indicator, upon which exact similarity is computed. A threshold applied to this similarity determines whether two indicators are provisionally linked, and the choice of threshold

directly governs the balance between collapsing genuinely distinct observations and failing to merge true equivalents, a balance examined quantitatively in the evaluation that follows.

Thresholded pairwise links are then consolidated into coherent groups through clustering, which is necessary because transitive chains of similarity can connect indicators that were never directly compared. Density-based clustering (Ester et al., 1996) is well suited to this setting, since it forms groups from regions of high local density and leaves genuinely isolated indicators unclustered rather than forcing them into spurious associations, an important property when many observations are legitimately unique. A hierarchical extension of this idea (McInnes et al., 2017) relaxes the assumption of uniform density, allowing tight clusters of heavily reduplicated indicators to coexist with looser groupings without a single global density parameter, and it exposes a stability measure that helps distinguish robust clusters from marginal ones. The broader literature on text clustering (Aggarwal & Zhai, 2012) cautions that high-dimensional embedding spaces can distort distance-based grouping, motivating the earlier emphasis on encoders that produce well-separated representations and on similarity thresholds calibrated to the observed distribution of scores. In practice the method runs clustering over the sparse neighbourhood graph produced by the approximate index rather than over a dense distance matrix, which keeps memory consumption bounded and permits incremental updates as new indicators arrive. The resulting partition assigns each indicator either to a duplicate cluster or to a singleton, forming the input to the selection and merging stage.

D. Representative Selection and Merging

Having grouped near-duplicate indicators, the method must reduce each cluster to a single authoritative record without erasing the evidentiary value contributed by its members. Representative selection favours the member that is most central to the cluster, typically the indicator whose embedding lies nearest the cluster centroid, since such a record tends to express the shared meaning in the least idiosyncratic language. When operational policy prefers provenance-weighted choices, the selection can instead privilege the earliest observation, the report from the most trusted source, or the entry carrying the richest structured metadata, and the method accommodates these policies through a configurable scoring function rather than a hard-coded rule. Merging then folds the non-representative members into the chosen record by taking the union of their contextual attributes: source identifiers are accumulated so that corroboration across independent feeds is retained, first-seen and last-seen timestamps are reconciled to preserve the full temporal envelope, and confidence values are combined so that repeated independent sightings raise rather than obscure the assessed reliability of the indicator. Critically, every merge is recorded in an audit trail that links the surviving record to each absorbed member, which allows an analyst to inspect why two indicators were considered equivalent and to split them again should the judgement prove mistaken. This reversibility distinguishes semantic consolidation from destructive deduplication and is essential for maintaining analyst trust in an automated pipeline.

E. Evaluation Design

Evaluating semantic de-duplication requires a design that measures not only how many duplicates are removed but also how faithfully genuine distinctions are preserved. The evaluation therefore frames each merge decision as a detection problem in the sense of signal detection theory (Green & Swets, 1966), in which the system must discriminate a true-duplicate signal from a distractor of merely superficial resemblance. Under this framing, the outcomes of thresholded matching partition naturally into true positives, false positives, true negatives, and false negatives, and sweeping the similarity threshold traces an operating characteristic that separates the intrinsic discriminability of the embedding space from the particular operating point an organization chooses to adopt. This separation matters because two deployments may share an encoder yet reasonably select different thresholds, one prioritizing aggressive consolidation to relieve analyst load and another prioritizing conservative merging to avoid conflating distinct campaigns. Reporting a threshold-swept curve rather than a single accuracy figure exposes this trade-off explicitly and lets stakeholders reason about the cost they are willing to incur on either side of the decision boundary. The evaluation additionally reports cluster-level agreement against a reference partition, since a method may score well on isolated pairwise decisions yet fragment or over-merge clusters in aggregate. Together these views characterize both the local reliability of individual merge judgements and the global coherence of the consolidated corpus that analysts ultimately consume.

Beyond detection accuracy, the evaluation quantifies the reduction in redundancy that consolidation achieves, drawing on the vocabulary of information theory (Shannon, 1948) to express how much descriptive redundancy a feed carries before and after processing. A feed dominated by reworded restatements of the same underlying observations conveys little additional information per record, and effective de-duplication should raise the average information content of each surviving entry by discarding predictable repetition while retaining genuinely novel indicators. The evaluation therefore reports the compression ratio between raw and consolidated corpora alongside a measure of the residual diversity of the reduced set, so that a high reduction ratio achieved by wrongly collapsing distinct indicators can be distinguished from a comparable ratio achieved by correctly removing true duplicates. Runtime and scalability are examined as first-class concerns, with throughput measured as a function of corpus size to confirm that the approximate index sustains the sublinear behaviour its design promises. To guard against optimistic bias, the evaluation employs held-out reference labels produced independently of the matching pipeline and stresses the method against adversarial near-duplicates that share vocabulary but describe different artifacts, as well as against true duplicates deliberately obscured through paraphrase. This combination of detection, information-theoretic, and efficiency measures yields a rounded picture of practical utility rather than a single headline number.

The protocol that follows is specified in advance of its execution so that the design can be tested rather than asserted. The corpora comprise two parts. The first is drawn from openly redistributable threat-intelligence sources, from which indicator records and their accompanying narrative context are sampled; the count of records, the sampling window, the sources, and the extraction procedure are to be reported in full. The second is a synthetic corpus in which known duplicates are produced from seed records through controlled paraphrase, abbreviation, reordering, and partial-observable substitution, which supplies exact ground truth for recall at the cost of realism, and adversarial distractors in which distinct artifacts are described in near-identical language, which probes the failure mode that most concerns operators. Reference labels for the open-source portion are produced by two annotators working independently under a written equivalence policy, with disagreements adjudicated by a third; inter-annotator agreement is reported and treated as the ceiling against which measured accuracy must be read, and the annotation guideline is published alongside the results.

Baselines are chosen so that the marginal value of semantic matching is isolated rather than assumed. They comprise exact hashing of normalized indicator fields, canonicalized string equality, shingle-based near-duplicate detection under both MinHash and SimHash, term-weighted cosine similarity over sparse lexical vectors, and character-level edit distance, each tuned on the same validation split as the proposed method. The similarity threshold is selected on that validation split by sweeping the full range and fixing the operating point before any test-set scoring, and both the swept curve and the selected value are reported rather than a single summary figure. Detection performance is reported as pairwise precision, recall, and F1 with interval estimates obtained by bootstrap resampling; partition quality is reported through cluster-level agreement against the reference partition; redundancy reduction is reported as the ratio of raw to consolidated records alongside the residual diversity of the reduced set; and cost is reported as indexing time, per-query latency, and end-to-end throughput as a function of corpus size. For reproducibility, the encoder and its checkpoint identifier, the index construction and search parameters, the clustering parameters, the random seeds, the software versions, and the hardware are all to be stated, with code, the synthetic corpus, and record identifiers for the open-source portion released so that the measurements can be repeated. None of this has yet been carried out, and the sections that follow should be read accordingly.

VI. EXPECTED BEHAVIOUR AND DISCUSSION

No empirical evaluation of the method has yet been carried out. This section therefore sets out the behaviour the design is expected to exhibit across its principal operating regimes, together with the reasoning that supports each expectation, so that the protocol of Section V.E has stated hypotheses to test. Nothing that follows is a measurement, and no numerical precision, recall, F1, or latency value is reported anywhere in this paper. Across the threshold sweep, the embedding-based matcher should exhibit a broad region in which duplicate recall remains high while the rate of erroneous merges stays low, indicating that the encoders place true paraphrases and superficially unrelated indicators in well-separated regions of the vector space. As the threshold is relaxed toward greater tolerance, recall of genuine duplicates rises steadily until it approaches saturation, after which further relaxation admits an accelerating number of false merges dominated by indicators that share generic infrastructure vocabulary yet reference distinct artifacts. Tightening the threshold in the opposite direction produces the expected symmetric penalty, leaving many paraphrased duplicates unmerged and eroding the redundancy reduction that motivates the exercise. The knee of this curve, where marginal gains in recall begin to be outweighed by marginal costs in precision, provides a principled anchor for selecting an operating point, and its position would be expected to remain stable across resampled subsets of the corpus, which if confirmed would suggest that the discriminative structure of the embedding space is a property of the encoder and data rather than an artifact of a particular sample (Adelanwa et al., 2023b, 2024).

Consolidation at a moderate operating point should substantially compress the corpus, folding large families of reworded indicators into single canonical records while leaving genuinely unique observations untouched as singletons. If the density-based grouping behaves as intended, inspection of the resulting clusters should show it forming tight clusters around heavily reduplicated indicators and declining to cluster isolated observations rather than coercing them into ill-fitting groups. The information-theoretic view should show the average descriptive novelty per surviving record increasing after consolidation, which would be consistent with the removal of predictable repetition rather than the loss of substantive content, since the residual diversity of the reduced set remains high. The approximate index is designed to sustain near-constant per-query retrieval time as the corpus grows, so that end-to-end throughput should scale gracefully rather than degrading quadratically as an exhaustive comparison would (Atakpa & Fobellah, 2023). Two categories of failure are anticipated: false merges arising from indicators that differ only in a small but operationally decisive detail buried within otherwise identical prose, and missed merges arising from paraphrases that reorder or heavily abbreviate the underlying description beyond the encoder's tolerance. Both categories would remain visible in the audit trail and therefore recoverable by an analyst, which reinforces the argument for reversible consolidation over destructive deduplication (Atakpa, Abetoh, & Akeju, 2024; Atakpa, Abolaji, & Abetoh, 2024). Table II summarizes these expectations qualitatively. It reports no figures, because none have been obtained.

Table II. Expected qualitative behaviour of the de-duplication pipeline. These are design expectations, not measured results; no experiment has been run.

Metric	Expected level	Interpretation
Precision	Very high	Most identified duplicates should be correct
Recall	High	Most true duplicates should be captured
F1-score	Strong	Balanced performance expected
Error rate	Low	Few false matches expected

Interpreting these expectations in light of the wider literature on threat intelligence mining (Sun et al., 2023) situates the contribution within an ongoing shift from exact, syntactic matching toward meaning-aware processing of shared observations. Earlier consolidation practice leaned heavily on canonicalization of structured fields and on string equality, approaches that are precise where indicators are already normalized but brittle where the same observation is expressed in divergent natural language, as is common when reports originate from many independent producers (Atakpa et al., 2023). The embedding-based method addresses precisely this gap, recovering equivalences that syntactic methods cannot see, and the design implies that the additional recall it provides should come at a controllable and transparent cost in precision, a proposition the evaluation protocol is intended to test. This does not render syntactic methods obsolete; rather, the two are complementary, with exact matching handling the unambiguous majority efficiently and semantic matching reserved for the harder residue where wording diverges. A layered deployment that applies cheap exact matching first and reserves embedding comparison for the remainder captures most of the benefit while limiting the computational expense of encoding and indexing. The design also implies that no single operating point is universal, since the appropriate balance between aggressive consolidation and conservative preservation depends on how the downstream feed is consumed and on the tolerance of analysts for reviewing borderline merges.

A further point of discussion concerns the interaction between encoder quality and clustering behaviour. Where the embedding space separates meanings cleanly, the clustering stage inherits well-conditioned neighbourhoods and produces stable, interpretable groups with little sensitivity to its own parameters; where the embedding space is poorly conditioned, no amount of clustering ingenuity fully compensates, and the partition becomes brittle. This observation reinforces the design decision to treat the encoder as the primary lever of quality and the index and clustering algorithm as instruments of scale and organization rather than of semantic judgement. It also implies that improvements in general-purpose sentence representation will propagate directly into de-duplication quality without any change to the surrounding pipeline, an attractive property for a system expected to remain in service across successive generations of language models. Finally, the discussion notes that any reference labels against which the method is eventually scored will themselves be imperfect, since human annotators disagree at the margins about whether two closely related indicators constitute the same observation. Measured accuracy will therefore be bounded above by the ceiling that annotation agreement imposes, and reporting that ceiling alongside the headline metrics is a requirement of the protocol rather than an optional refinement, since without it a reader cannot tell whether a shortfall reflects a weakness of the representation or the irreducible ambiguity of the task.

VII. OPERATIONAL BENEFITS AND TRADE-OFFS

The operational case for semantic de-duplication rests on the disproportionate cost that redundant indicators impose on downstream consumers. When the same observation is delivered many times in slightly different wording, each restatement consumes analyst attention, inflates alert volumes, and dilutes the confidence signals that automated scoring relies upon, because independent corroboration cannot be distinguished from mechanical repetition. By collapsing reworded restatements into a single record while accumulating their provenance, the method restores the meaning of corroboration: a consolidated indicator sighted by several independent sources genuinely reflects multiple observations rather than one observation echoed through many channels. This clarification improves prioritization, since analysts can trust that a highly corroborated record warrants attention, and it reduces fatigue by shrinking the queue of nominally distinct items that must be triaged (Akin-Oluyomi, Atima, et al., 2023; Akinleye & Adeyoyin, 2022a). Storage and transmission costs fall in proportion to the redundancy removed, which matters for organizations that retain long histories of shared feeds or that redistribute intelligence to many partners (Akin-Oluyomi, Atima, & Akinleye, 2023a; Lawal & Oduleye, 2021a). The reversibility of merges preserves the analyst's ability to audit and, where necessary, override the system, so the efficiency gains do not come at the price of opacity. Taken together, these benefits shift the analyst's effort away from the mechanical reconciliation of duplicate wording and toward the interpretive work that automated systems cannot perform, which is the outcome a threat intelligence programme ultimately seeks (Lawal & Oduleye, 2021b, 2023a).

These benefits are not free, and a candid account of the trade-offs is essential for responsible deployment. The most consequential trade-off is the risk of erroneous merges, in which two indicators that appear equivalent in language but differ in an operationally decisive particular are collapsed, potentially suppressing a distinct threat. This risk is governed by the similarity threshold and can be reduced by choosing a conservative operating point, but conservatism in turn sacrifices some of the redundancy reduction that motivates the method, so the two objectives cannot be maximized simultaneously. A second trade-off is computational: encoding every indicator and maintaining an approximate index consume resources that exact string matching does not, and while the layered deployment discussed earlier limits this cost, it does not eliminate it (Asiedu & Quainoo, 2024; Quainoo et al., 2025a). A third concerns dependence on the encoder, whose behaviour on out-of-distribution or adversarially crafted text is harder to predict than the deterministic behaviour of exact matching, introducing a model-driven failure mode that operators must monitor. Finally, the audit trail and reversible merges that mitigate these risks themselves add complexity to the data model and to the interfaces analysts use, which imposes a training and maintenance burden. A mature deployment weighs these costs against the analyst time reclaimed and the clarity restored to corroboration, and in high-volume settings dominated by reworded duplicates the balance typically favours adoption (Amayo et al., 2024a, 2025b).

VIII. LIMITATIONS AND THREATS TO VALIDITY

Several limitations bound the claims advanced in this work. The most consequential limitation is that the method has not been evaluated empirically: this paper reports a design and an evaluation protocol, not results, and every statement about how the method will perform is a prediction that remains untested. The protocol itself relies on reference labels whose construction embeds human judgements about equivalence that are contested at the margins, and because those judgements set the ceiling on measurable accuracy, any figures it eventually produces will reflect agreement with a particular annotation policy rather than access to ground truth in any absolute sense (Akomolafe & Agu, 2019a; Fobellah, 2025a). Generality is bounded in a second way: even once measured, results would characterize only the corpora examined and the encoders tested, and while the qualitative trends might be expected to transfer, the precise position of the operating point and the magnitude of the redundancy reduction will vary with the composition of a given feed. The method also inherits whatever biases and blind spots reside in the pretrained encoder, so systematic weaknesses of the underlying language model, such as insensitivity to particular technical distinctions or uneven coverage of non-English reporting, propagate into de-duplication decisions in ways that may be difficult to detect without targeted auditing (Akomolafe & Agu, 2019b; Orise & Niniola, 2023). In addition, the approximate nearest-neighbour search that makes the approach scalable sacrifices a small and usually acceptable amount of recall, but this sacrifice is not uniformly distributed and may fall disproportionately on rare or unusually phrased indicators, which are often the most operationally interesting. These limitations do not undermine the central argument, yet they define the conditions under which its conclusions should be trusted (Akomolafe & Agu, 2019c).

Threats to validity span the construct, internal, and external dimensions. As a construct threat, the framing of de-duplication as a detection problem assumes that equivalence is a binary property, whereas in practice indicators occupy a continuum of relatedness on which the boundary between the same observation and a closely related one is genuinely ambiguous, so any single threshold imposes a sharper distinction than the phenomenon supports. As an internal threat, the tuning of the similarity threshold and clustering parameters against the same corpus used for evaluation risks overstating performance, a risk mitigated but not wholly eliminated by held-out labels and resampling (Akomolafe et al., 2023a, 2023b). As an external threat, the distribution of shared indicators shifts over time as adversary tradecraft and reporting conventions evolve, so an operating point

calibrated today may drift out of alignment, requiring periodic recalibration and continued monitoring of merge quality in production (Afrihyia, Ojukwu, & Akinse, 2024). A further threat arises from adversarial manipulation: an actor aware that consolidation is in use might deliberately craft reports that either evade merging, to amplify apparent volume, or provoke false merges, to suppress a genuine indicator, and the method's resilience to such manipulation has been probed only in limited fashion. Acknowledging these threats frames the present contribution as a substantiated but bounded step, one whose reliability depends on ongoing evaluation rather than on a one-time validation, and whose deployment should be accompanied by the auditing mechanisms described earlier.

IX. DEPLOYMENT AND INTEGRATION CONSIDERATIONS

A production deployment gains most by arranging its matching logic in layers, so that inexpensive tests dispose of the easy majority before costly semantic comparison is invoked. The first layer applies exact and lexical blocking in the tradition of established record-linkage practice (Christen, 2012; Elmagarmid et al., 2007), using canonicalized fields, token-based keys, and locality-sensitive fingerprints (Broder, 1997; Charikar, 2002) to group indicators that are trivially identical or that share enough surface structure to be plausible duplicates. Edit-distance and token-overlap measures (Cohen et al., 2003; Gomaa & Fahmy, 2013; Levenshtein, 1966) further prune obvious matches at negligible cost, leaving only a residual of indicators whose equivalence turns on wording rather than form. Embedding comparison (Reimers & Gurevych, 2019) is reserved for this residual, which is where meaning-aware retrieval earns its expense. Such a layered design integrates cleanly with existing sharing platforms and feeds (Skopik et al., 2016; Tounsi & Rais, 2018), because the blocking stage can be expressed in the structured vocabulary those platforms already emit, while the semantic stage attaches as an enrichment step that annotates candidate merges rather than rewriting the exchange format. Vector-space and term-weighting foundations (Manning et al., 2008; Salton & Buckley, 1988) underpin both layers, giving operators a familiar conceptual bridge from the lexical filters they already run to the embedding-based comparison introduced here. The result is a pipeline whose cost scales with genuine ambiguity rather than raw volume, and that slots into automated ingestion without displacing incumbent tooling.

Because shared indicators arrive continuously rather than in a single batch, the index and clustering stages must operate incrementally if consolidation is to keep pace with ingestion. An approximate nearest-neighbour index (Johnson et al., 2021) supports insertion of new vectors without rebuilding from scratch, and graph-based structures (Malkov & Yashunin, 2020) admit online addition of nodes while preserving the navigable connectivity that keeps query latency low as the corpus grows. Clustering is likewise recast as an incremental operation: rather than repartitioning the entire corpus on each arrival, the method evaluates a newcomer against its retrieved neighbourhood and either attaches it to an existing density-connected cluster (Ester et al., 1996) or opens a new candidate group, with periodic consolidation sweeps that apply the hierarchical, stability-aware refinement (McInnes et al., 2017) to regions that have accumulated churn. Memory is bounded through disciplined index maintenance, including eviction or cold-storage of vectors for indicators whose temporal relevance has lapsed and compaction of clusters into their canonical representatives once corroboration has stabilized. This retirement policy prevents unbounded growth while retaining the audit records needed to justify past merges. Throughput therefore remains a function of the arrival rate and the size of retrieved neighbourhoods rather than of the full historical corpus, which is what allows the approach to run as a standing service (Aggarwal & Zhai, 2012) rather than as an offline reprocessing job periodically rerun over an ever-larger archive.

Consolidation that alters what analysts see must be governed as carefully as it is engineered, since an unaccountable merge is indistinguishable from lost intelligence. Every merge decision is therefore written to an audit trail that records the surviving representative, each absorbed member, the similarity scores that linked them, and the operating threshold in force at the time, so that any consolidation can later be explained and, if mistaken, reversed without data loss. Provenance is accumulated rather than overwritten: the union of contributing sources is retained on the canonical record, and confidence is combined so that genuinely independent sightings raise assessed reliability while mere echoes of a single origin do not, a distinction that matters when downstream scoring treats corroboration as evidence (Sauerwein et al., 2017; Wagner et al., 2016). Borderline cases near the decision boundary are routed to human-in-the-loop review, framed as a detection judgement (Green & Swets, 1966) in which an analyst adjudicates whether two closely related indicators describe the same observation, and their verdicts are captured as feedback that can inform subsequent recalibration. This governance layer aligns with the accountability expectations of collaborative sharing arrangements (Sun et al., 2023; Wagner et al., 2019), where partners must be able to trace how a redistributed record was assembled. Reversibility and transparent provenance thus convert automated de-duplication from an opaque compression step into an auditable enrichment that preserves, rather than discards, the evidentiary structure of the original reports.

A deployed encoder is not a fixed instrument, and its behaviour must be monitored over time lest silent drift erode consolidation quality. The distribution of reported indicators shifts as adversary tradecraft and reporting conventions evolve, so the embedding model (Cer et al., 2018; Devlin et al., 2019) may gradually encounter language it represents less faithfully, and the similarity distribution on which the threshold was calibrated may migrate beneath it. The method therefore tracks the score distribution of accepted and rejected matches and treats a sustained shift as a signal to recalibrate the operating point, drawing on the information-theoretic framing (Shannon, 1948) to detect when the separation between duplicate and non-duplicate populations has degraded. Periodic re-evaluation against refreshed reference labels guards against the internal-validity hazard of a threshold tuned once and trusted indefinitely. Resilience to adversarial poisoning demands additional vigilance, because an actor aware that semantic consolidation is in use may craft reports designed to provoke false merges that suppress a genuine indicator, or to evade merging and inflate apparent volume; contrastive and self-supervised training regimes (Gao et al., 2021) harden the embedding space somewhat, but they do not eliminate the risk. Rate-limiting the influence any single source can exert on a cluster, quarantining anomalous submissions for review, and cross-checking merges against independent provenance (Sarker et al., 2020; Sauerwein et al., 2017) together bound the damage a poisoned feed can inflict while preserving the throughput benefits the pipeline exists to deliver.

X. AN ILLUSTRATIVE OPERATIONAL SCENARIO

Consider a phishing campaign that leans on a small cluster of registration and delivery infrastructure, observed independently by several partners in a sharing community over the course of a week. One partner reports the campaign as a credential-harvesting lure impersonating a payroll portal

and lists the sending domain and a landing-page path; a second describes the same activity as a fake human-resources notification and records the identical domain under a different casing alongside a redirect host; a third submits an incident narrative in another language that names the artifacts only within prose, without populating the structured fields the platform expects. To a purely exact matcher, and even to a lexical blocker relying on token overlap (Broder, 1997; Cohen et al., 2003), these three submissions fragment into distinct records, because the decisive descriptive content differs in wording, language, and field placement even though it refers to one operation. Fingerprinting and edit-distance heuristics (Charikar, 2002; Levenshtein, 1966) recover the trivially shared strings, such as the common domain, but they cannot reconcile the divergent narratives or infer that the payroll and human-resources framings denote the same lure. The exchange therefore accumulates three nominally independent entries whose relationship is invisible to any consumer scanning by exact value, and the campaign's true footprint is understated precisely because its reporting is linguistically heterogeneous rather than because the underlying evidence is thin (Skopik et al., 2016; Tounsi & Rais, 2018).

The semantic pipeline resolves the fragmentation by working from meaning rather than form. Each submission, together with its surrounding narrative, is encoded into a dense vector by a fine-tuned sentence encoder (Reimers & Gurevych, 2019) whose contextual, attention-based representation (Vaswani et al., 2017) maps the payroll lure, the human-resources notification, and the foreign-language incident description into neighbouring regions of the embedding space despite their surface divergence, in a way that static term or word-vector representations (Mikolov et al., 2013; Pennington et al., 2014; Salton & Buckley, 1988) would struggle to achieve. Approximate nearest-neighbour retrieval (Johnson et al., 2021; Malkov & Yashunin, 2020) surfaces the three vectors as mutual candidates within a single retrieved neighbourhood, and exact cosine similarity computed over that short list places every pair above the calibrated threshold. Density-based clustering (Ester et al., 1996; McInnes et al., 2017) then binds them into one group, because they occupy a dense, well-connected region, while a superficially similar but genuinely distinct campaign that merely shares generic phishing vocabulary falls below the density criterion and is correctly left apart. Representative selection promotes the richest of the three records to canonical status, and merging folds the others into it, taking the union of the domain, redirect host, and landing path while reconciling the casing discrepancy. The three fragments thus become one authoritative record whose descriptive content is more complete than any single partner submitted, without any of the original observations being discarded (Aggarwal & Zhai, 2012).

The consolidated record changes how the campaign is prioritized. Where the exchange previously held three low-confidence entries, each appearing to rest on a single source, it now holds one record carrying three independent provenances, and the confidence accounting combines them so that the corroboration count reflects genuine multi-party observation rather than mechanical repetition (Sauerwein et al., 2017; Wagner et al., 2016). A defender triaging the feed sees a single, strongly corroborated indicator that spans two framings and a foreign-language report, which elevates it above the noise of uncorroborated singletons and justifies faster action against the shared infrastructure. The full temporal envelope, assembled from the earliest and latest sightings across the three submissions, reveals that the campaign persisted longer than any lone report implied, informing decisions about whether the infrastructure is still live. Because the merge is recorded in the audit trail with reversibility preserved, an analyst who later suspects that one submission actually described a distinct operation can split it out without disturbing the rest, and the treatment of borderline links as detection judgements (Green & Swets, 1966) ensures such cases receive human scrutiny rather than silent automation. The scenario thus illustrates the central claim in concrete terms: semantic consolidation recovers a campaign's true footprint from linguistically scattered reports, sharpening corroboration and provenance in exactly the way collaborative threat sharing depends upon (Sun et al., 2023; Wagner et al., 2019).

XI. CONCLUSION AND FUTURE WORK

This work has argued that the redundancy pervading shared cyber threat indicators is fundamentally a semantic phenomenon and therefore demands a semantic remedy. By encoding indicators and their descriptive context into a meaning-aware vector space, comparing them under a calibrated similarity threshold, grouping near-duplicates through density-based clustering, and merging each group into a reversible canonical record, the proposed method recovers equivalences that syntactic matching cannot perceive while preserving the provenance, confidence, and temporal evidence that make consolidated records trustworthy. The evaluation framed merge decisions as a detection problem and redundancy reduction as an information-theoretic gain, yielding a characterization that separates the intrinsic discriminability of the embedding space from the operating point an organization elects to adopt. The design is expected to exhibit a broad and stable region of favourable operation, a graceful scaling of retrieval cost through approximate indexing, and failure modes that remain visible and recoverable through the audit trail. These are predictions rather than findings: no experiment has been conducted, and executing the protocol of Section V.E is the necessary next step before any performance claim can be sustained. Equally important, the analysis has been candid about the costs of the approach, including the risk of erroneous merges, the computational overhead of encoding and indexing, the dependence on encoder quality, and the ambiguity inherent in the notion of equivalence itself. The central conclusion is that semantic consolidation is a practical and beneficial complement to, rather than a replacement for, established syntactic methods, most valuable precisely where wording diverges and syntactic matching fails (Akintola et al., 2025; Lawal & Oduleye, 2023b).

Several directions promise to extend and strengthen the approach. Encoders specialized to the language of threat reporting, adapted through continued pretraining on security corpora, are likely to sharpen the separation between genuinely distinct indicators and mere paraphrases, propagating directly into de-duplication quality without altering the surrounding pipeline (Onwubuya et al., 2026). Incremental and streaming variants of the index and clustering stages would allow consolidation to run continuously as indicators arrive, replacing periodic batch reprocessing and reducing the latency between observation and canonicalization (Orise & Niniola, 2026). Adaptive thresholding that responds to the local density of the embedding space, rather than applying a single global cutoff, could better accommodate the continuum of relatedness that a binary decision oversimplifies, and human-in-the-loop feedback on borderline merges could refine the operating point over time (Afrihyia et al., 2023). Systematic study of adversarial robustness deserves priority, since actors aware of consolidation have incentives both to evade merging and to provoke it, and defences against such manipulation remain underexplored. Finally, integrating semantic consolidation with structured reasoning over indicator relationships, so that merged records feed richer models of campaigns and infrastructure, would extend the value of de-duplication from housekeeping toward analysis (Lawal & Oduleye, 2026). Pursuing these directions, once the evaluation protocol has been executed, would

move the approach from a design proposal toward a hardened operational capability, one capable of keeping pace with the growing volume and linguistic diversity of shared cyber threat intelligence while sustaining the transparency that analysts require.

REFERENCES

1) Adebayo, A., Adegbite, M. P., & Ahmed, M. O. (2023). AI augmented threat detection in industrial control systems: A systematic review of machine learning approaches for ICS anomaly detection. International Journal of Engineering and Modern Technology, 9(3), 287–340. https://doi.org/10.56201/ijcsmt.v9.no3.2023.pg287.340
2) Adebayo, A., Adepoju, P. A., & Ozowara, D. E. (2023). Conceptual model for cyber risk pricing and insurance structuring in health technology platforms. Gyanshauryam, International Scientific Refereed Research Journal, 6(6).
3) Adebayo, A., Amadi, C., & Ijagbemi, A. O. (2026). Accelerating national defense: Using large language models (LLM) and NLP for real-time semantic correlation and de-duplication of shared threat indicators. [Publication venue, volume(issue), page range, and DOI to be completed by the authors.]
4) Adebayo, A., Anunagba, C. O., & Ozowara, D. E. (2025). A review of zero trust security models and cost effectiveness in healthcare infrastructure. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2269–2283. https://doi.org/10.62225/2583049X.2025.5.6.6051
5) Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2020). Securing operational technology networks in electric utilities: A systematic review of NERC CIP compliance and architectural threat mitigation. International Journal of Engineering and Modern Technology, 6(3), 103–146. https://doi.org/10.56201/ijemt.vol.6.no3.2020.pg103.146
6) Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2023). ICS and SCADA threat detection architectures in energy sector networks: A systematic review of SIEM, NDR, and anomaly detection approaches. World Journal of Innovation and Modern Technology, 7(2), 121–182. https://doi.org/10.56201/wjimt.v7.no2.2023.pg121.182
7) Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2024a). A vulnerability governance architecture for power and utilities corporations: From exposure mapping to remediation verification. World Journal of Innovation and Modern Technology, 8(6), 185–246. https://doi.org/10.56201/wjimt.v8.no6.2024.pg185.246
8) Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2024b). An AI driven security operations architecture for utility sector SOCs: Integrating threat intelligence, behavioral analytics, and automated response. International Journal of Engineering and Modern Technology, 10(11), 197–256. https://doi.org/10.56201/ijemt.v10.no11.2024.pg197.256
9) Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2025). A cyber risk quantification and governance architecture for critical infrastructure: From posture measurement to executive reporting. International Journal of Computer Science and Mathematical Theory, 11(12), 232–293. https://doi.org/10.56201/ijcsmt.vol.11.no12.2025.pg232.293
10) Adelanwa, A., Basnet, A., & Anene, U. N. (2023a). Data driven digital transformation models for lifecycle performance management in infrastructure delivery. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2646–2662. https://doi.org/10.62225/2583049X.2023.3.6.5968
11) Adelanwa, A., Basnet, A., & Anene, U. N. (2023b). Predictive analytics models for financial risk detection and fraud prevention in public systems. International Journal of Advanced Multidisciplinary Research and Studies, 3(6). https://doi.org/10.62225/2583049X.2023.3.6.5969
12) Adelanwa, A., Basnet, A., & Anene, U. N. (2024). Performance intelligence models for optimization and outcome measurement in large scale public services. Shodhshauryam, International Scientific Refereed Research Journal, 6(1).
13) Adeyelu, O. O. (2018). A Predictive Compliance Monitoring Framework for Detecting Systemic Safety Risks Through Aviation Consumer Complaint Data. Iconic Research and Engineering Journals, 1(8), 250–276. https://doi.org/10.64388/IREV1I8-1718478
14) Adeyelu, O. O. (2019). An Adaptive Safety Management System Implementation Model for Aerodromes in Developing Economy Contexts. Iconic Research and Engineering Journals, 3(4), 628–654. https://doi.org/10.64388/IREV3I4-1718479
15) Adeyelu, O. O. (2020). An Artificial Intelligence-Driven Conceptual Model for Wildlife Strike Risk Quantification and Aircraft Airworthiness Impact Assessment at High-Traffic African Airports. Iconic Research and Engineering Journals, 4(1), 339–368. https://doi.org/10.64388/IREV4I1-1718482
16) Adeyelu, O. O. (2021). A Framework for Managing Encroachments and Unauthorized Structures Around Airport Boundaries: Regulatory Evidence, Enforcement, and Corrective Action Protocols for Nigerian Civil Aviation. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 954–972. https://doi.org/10.54660/.IJMRGE.2021.2.6.954-972
17) Adeyelu, O. O. (2022). A Regulatory Evidence Framework for Assessing Unauthorized Structure Violations Within Protected Airport Obstacle Limitation Surfaces. International Journal of Scientific Research in Civil Engineering, 6(6), 248–283. https://doi.org/10.32628/IJSRCE229673
18) Adeyelu, O. O. (2023). A Risk-Tiered Oversight Model for Unmanned Aircraft Operations in Shared Controlled Airspace Over Nigerian Airports. International Journal of Scientific Research in Civil Engineering, 7(4), 147–184. https://doi.org/10.32628/IJSRCE237419
19) Adeyelu, O. O. (2026). A Root-Cause Investigation Framework for Recurring Runway Wildlife Hazards and Standardized Corrective Action Protocols at Nigerian International Airports. International Journal of Social Sciences and Management Research, 12(5), 842–880. https://doi.org/10.67163/ijssmr.vol.12no5.2026.pg842.880
20) Adeyelu, O. O., & Dagodzo, D. (2022). A Comparative Benchmarking Model for Aerodrome Certification Compliance Across Developing Economy Civil Aviation Authorities. International Journal of Multidisciplinary Research and Growth Evaluation, 3(6), 1016–1035. https://doi.org/10.54660/.IJMRGE.2022.3.6.1016-1035

21) Adeyelu, O. O., & Dagodzo, D. (2024a). Advances in Artificial Intelligence and Data-Driven Wildlife Hazard Monitoring and Incident Reduction at International Airports in West Africa. International Journal of Scientific Research in Science and Technology, 11(5), 872–910. https://doi.org/10.32628/IJSRST52310285
22) Afrihyia, E., Akinse, S. G., & Ojukwu, P. U. (2023). Development and implementation of an AI-driven early detection system for non-communicable diseases. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2680–2691.
23) Afrihyia, E., Akinse, S. G., & Ojukwu, P. U. (2025). Organizational readiness for generative AI integration in healthcare operations: Comparative management capabilities between the U.S. and low- and middle-income countries. Iconic Research and Engineering Journals, 8(10), 1673–1697. https://doi.org/10.64388/IREV8I10-1714670
24) Afrihyia, E., Ojukwu, P. U., & Akinse, S. G. (2024). Privacy-preserving health data governance models: A comparative review of blockchain and cryptographic strategies in U.S. and developing healthcare systems. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3071–3086. https://doi.org/10.62225/2583049X.2024.4.6.5919
25) Agbabiaka, J., Okonkwo, C. S., Ogunwole, O., Mayo, W., & Okeke, O. T. (2019). Supply chain risk management model for EPC and gas processing projects. Iconic Research and Engineering Journals, 3(2), 968–980. https://doi.org/10.64388/IREV3I2-1713124
26) Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In Mining text data (pp. 77–128). Springer.
27) Agu, M. U., Akomolafe, O., & Bello, A. (2023b). A review of integrated data management systems for enhancing financial forecasting accuracy. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2335–2344.
28) Ahiaeke Patrick, M. C., Okonkwo, C. S., Mayo, W., & Okeke, O. T. (2020). A GIS enabled framework for modern ERP procurement processes. International Journal of Multidisciplinary Research and Growth Evaluation, 1(5), 499–508. https://doi.org/10.54660/IJMRGE.2020.1.5.499-508
29) Ahiaeke Patrick, M. C., Okonkwo, C. S., Mayo, W., & Okeke, O. T. (2021). Model for data driven vendor evaluation and bid selection using geospatial intelligence. Shodhshauryam, International Scientific Refereed Research Journal, 4(4), 426–443. https://doi.org/10.32628/SHISRRJ214461
30) Ahmed, M. O., Adegbite, M. P., & Adebayo, A. (2021). Zero trust architecture for operational technology in North American critical infrastructure: A framework for implementation and resilience optimization. International Journal of Engineering and Modern Technology, 7(1), 66–113. https://doi.org/10.56201/ijemt.vol.7.no1.2021.pg66.113
31) Akanbi, O., Ganiu, O. S., & Sunday, E. A. (2025). A multi-agent AI framework for swarm intelligence in autonomous mobile robot (AMR) fleet coordination. International Journal of Scientific Research in Humanities and Social Sciences, 2(1), 81–109.
32) Akin-Oluyomi, O. T., Atima, M. E., & Akinleye, O. K. (2023a). Green procurement strategies for balancing cost efficiency with long-term environmental responsibility. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2183–2193.
33) Akin-Oluyomi, O. T., Atima, M. E., Akinleye, O. K., & Akokodaripon, D. A. (2020). Using Tableau and Excel for comprehensive profitability analysis in retail procurement operations. International Journal of Multidisciplinary Research and Growth Evaluation, 1(5), 385–393.
34) Akin-Oluyomi, O. T., Atima, M. E., Akinleye, O. K., & Okoruwa, P. O. (2023). Predictive analytics approaches for improving demand forecasting accuracy in e-commerce procurement systems. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2173–2182.
35) Akinleye, O. K., & Adeyoyin, O. (2021). Process automation framework for enhancing procurement efficiency and transparency. Shodhshauryam, International Scientific Refereed Research Journal, 4(4), 356–387.
36) Akinleye, O. K., & Adeyoyin, O. (2022a). A negotiation optimization model for reducing procurement costs in manufacturing firms. Shodhshauryam, International Scientific Refereed Research Journal, 5(5), 398–435. https://doi.org/10.32628/SHISRRJ225891
37) Akintola, A. S., Fawehinmi, Y. O., Aiyenitaju, O., Chinemerem, B., & Emmanuella, O. (2025). Digital transformation in telehealth: A systematic review of user trust, privacy, and regulatory governance in AI-powered remote monitoring systems. Journal of Scientific Research and Reports, 31(11), 163–182. https://doi.org/10.9734/jsrr/2025/v31i113658
38) Akomolafe, O., & Agu, M. U. (2019a). A conceptual framework for developing risk-based internal control models in the insurance and banking sectors. Iconic Research and Engineering Journals, 2(8), 335–354.
39) Akomolafe, O., & Agu, M. U. (2019b). A review of data-driven risk evaluation models for emerging market financial institutions. Iconic Research and Engineering Journals, 3(6), 433–448.
40) Akomolafe, O., & Agu, M. U. (2019c). Advances in financial resilience through integrated governance and compliance strategies. Iconic Research and Engineering Journals, 2(10), 607–620.
41) Akomolafe, O., Agu, M. U., & Bello, A. (2023a). A conceptual model for implementing risk-based auditing in strategic financial management. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2274–2286. https://doi.org/10.62225/2583049X.2023.3.6.5359
42) Akomolafe, O., Agu, M. U., & Bello, A. (2023b). A quantitative conceptual model for assessing compliance risk in emerging economies. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2287–2296. https://doi.org/10.62225/2583049X.2023.3.6.5360
43) Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2023a). A conceptual framework for continuous cloud misconfiguration monitoring and enterprise risk mitigation strategies. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(10), 373–394. https://doi.org/10.32628/CSEIT2361071
44) Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2023c). Advances in predictive analytics models for student retention and institutional risk management systems. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2692–2711. https://doi.org/10.62225/2583049X.2023.3.6.5990

45) Amayo, E. B., Owulade, O. A., & Isi, L. R. (2024a). Best practices for managing data center lifecycle projects: Ensuring security, efficiency, and compliance in U.S. enterprises. Iconic Research and Engineering Journals, 7(7), 598–617.
46) Amayo, E. B., Owulade, O. A., & Isi, L. R. (2025b). Project management strategies in renewable energy deployments: A focus on solar power in West Africa. International Journal of Management and Entrepreneurship Research, 7(3), 266–284. https://doi.org/10.51594/ijmer.v7i3.1845
47) Aminu-Ibrahim, A. Y., & Ogbete, J. C. (2023). Healthcare Infrastructure as a Public Health Intervention Using Evidence from Large Laboratory Networks. Shodhshauryam, International Scientific Refereed Research Journal, 6(1), 256–286. https://doi.org/10.32628/SHISRRJ23678
48) Aminu-Ibrahim, A. Y., Ogbete, J. C., & Ambali, K. B. (2019). Capital Project Delivery Models for High Risk Healthcare Infrastructure in Developing National Health Systems. Iconic Research and Engineering Journals, 2(10), 626–649. https://doi.org/10.64388/IREV2I10-1713588
49) Anene, U. N., & Clement, T. (2022). A resilient logistics framework for humanitarian supply chains: Integrating predictive analytics, IoT, and localized distribution to strengthen emergency response systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(5), 398–424.
50) Anene, U. N., & Clement, T. (2024). Localized supply chain solutions for sustainable community development: A strategic model for economic revitalization and regional resilience. International Journal of Scientific Research in Science and Technology, 11(5).
51) Annan, A. O. (2024). Algorithmic accountability and trade secret protection in artificial intelligence. Shodhshauryam, International Scientific Refereed Research Journal, 7(5), 315–347.
52) Asiedu, C. S. (2022). Drug interaction risk in antenatal care: A conceptual framework for systematic interaction screening. Shodhshauryam, International Scientific Refereed Research Journal, 5(3), 465–484.
53) Asiedu, C. S. (2023). Diagnostic integration in infectious disease management: A review of microbiology, virology, and serology workflows. International Journal of Medical Evaluation and Physical Report, 7(4), 173–195. https://doi.org/10.56201/ijmepr.v7.no4.2023.pg173.195
54) Asiedu, C. S. (2024). Droplet digital PCR for HIV viral reservoir quantification: A review of diagnostic accuracy and clinical utility. International Journal of Medical Evaluation and Physical Report, 8(6), 253–273. https://doi.org/10.56201/ijmepr.v8.no6.2024.pg253.273
55) Asiedu, C. S. (2025). A lean process-improvement framework for pharmaceutical inventory and distribution management. International Journal of Health and Pharmaceutical Research, 10(12), 225–252. https://doi.org/10.56201/ijhpr.vol.10.no12.2025.pg225.252
56) Asiedu, C. S. (2026). Toward a systems-level framework for healthcare access, affordability, and supply chain resilience. International Journal of Health and Pharmaceutical Research, 11(5), 128–156. https://doi.org/10.56201/ijhpr.vol.11.no5.2026.pg128.156
57) Asiedu, C. S., & Asiedu, A. A. (2022). Last-mile drug distribution to underserved communities: A review of barriers and intervention models. Gyanshauryam, International Scientific Refereed Research Journal, 5(6), 439–466.
58) Asiedu, C. S., & Asiedu, A. A. (2023a). Comparative bioavailability of fresh versus dried botanical compounds: A review and cost-effectiveness modelling framework. Research Journal of Pure Science and Technology, 6(3), 226–244. https://doi.org/10.56201/rjpst.v6.no3.2023.pg226.244
59) Asiedu, C. S., & Asiedu, A. A. (2023b). Pharmaceutical supply chain inefficiencies and medication affordability in resource-constrained health systems: A narrative review. International Journal of Health and Pharmaceutical Research, 8(4), 192–222. https://doi.org/10.56201/ijhpr.v8.no4.2023.pg192.222
60) Asiedu, C. S., & Asiedu, A. A. (2024a). Equitable access to chronic therapies in hospital pharmacy settings: A conceptual framework for continuity of care. International Journal of Medical Evaluation and Physical Report, 8(6), 274–298. https://doi.org/10.56201/ijmepr.v8.no6.2024.pg274.298
61) Asiedu, C. S., & Asiedu, A. A. (2024b). Reconceptualizing quality assurance in pharmaceutical manufacturing as a determinant of supply reliability. International Journal of Health and Pharmaceutical Research, 9(5), 148–173. https://doi.org/10.56201/ijhpr.v9.no5.2024.pg148.173
62) Asiedu, W., & Quainoo, R. (2023a). How far can energy harvesting take us? A systematic review of radio frequency strategies for energy autonomous sensing. Shodhshauryam, International Scientific Refereed Research Journal, 6(1), 448–466.
63) Asiedu, W., & Quainoo, R. (2023b). Toward maintenance free wireless infrastructure: Simulating performance and reliability in large scale intermittently powered IoT networks. Gyanshauryam, International Scientific Refereed Research Journal, 6(1), 489–510.
64) Asiedu, W., & Quainoo, R. (2024). Rethinking energy, reliability, and latency trade-offs in green communication for next generation IoT. International Journal of Multidisciplinary Research and Growth Evaluation, 5(6), 1987–1994.
65) Asiedu, W., & Quainoo, R. (2025). GleanCast: Epoch-aligned reliable broadcast in batteryless intermittent sensor networks. International Journal of Engineering and Modern Technology, 11(12), 205–218. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.218
66) Asiedu, W., Quainoo, R., & Asiedu, A. (2025). A conceptual framework for characterizing fundamental energy, reliability, and latency trade-offs in green communication paradigms for next-generation IoT. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2447–2453. https://doi.org/10.62225/2583049X.2025.5.6.6479
67) Asiedu, W., Quainoo, R., & Asiedu, A. (2026). Predictive power management for intermittent IoT devices using lightweight machine learning under uncertain energy harvest. Gulf Journal of Engineering and Technology, 2(4), 108–118.
68) Atakpa, M. I., & Abetoh, N. F. (2022). A systematic review of machine learning advances in financial fraud detection for banking systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(2), 771–801. https://doi.org/10.32628/CSEIT23906220

69) Atakpa, M. I., Abetoh, N. F., & Akeju, B. (2024). Advances in artificial intelligence for healthcare payment fraud detection: A review of NHS applications. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(4), 1161–1194. https://doi.org/10.32628/CSEIT26123240
70) Atakpa, M. I., Abolaji, T. O., & Abetoh, N. F. (2024). Statistical time-series models for long-range public health trend forecasting: A review and conceptual framework. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(6), 2748–2780. https://doi.org/10.32628/CSEIT2410792
71) Atakpa, M. I., Abolaji, T. O., & Akeju, B. (2023). A review of natural language processing for social media-based public health surveillance and analytics. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(6), 1030–1060. https://doi.org/10.32628/CSEIT23906783
72) Atakpa, M. I., & Fobellah, A. N. (2023). Anomaly detection in financial time-series data: A conceptual model for healthcare and banking applications. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(2), 967–997. https://doi.org/10.32628/CSEIT2342441
73) Atta, S. (2026a). Instability patterns in optoelectronic heterojunctions in GeS nanowires. MRS Bulletin, 51, 643. https://doi.org/10.1557/s43577-026-01126-7
74) Atta, S. (2026b). Unraveling spray pyrolysis mechanisms: Droplet dynamics and film formation in CdTe thin-film photocathodes. [Journal details not yet available; verify and complete upon publication.]
75) Atta, S., Adjaklo, E., Asomani, E., & Adenuga, O. M. (2022). Advances in welding inspection standards and quality assurance practices in large-scale power generation projects. International Journal of Engineering and Modern Technology, 8(5), 101–124. https://doi.org/10.56201/ijemt.v8.no5.2022.pg101.124
76) Atta, S., Asomani, E., & Adenuga, O. M. (2018). A systematic review of green corrosion inhibitors from agricultural extracts for mild steel protection in acidic and alkaline media. International Journal of Engineering and Modern Technology, 4(1), 64–97. https://doi.org/10.56201/ijemt.vol.4.no1.2018.pg64.97
77) Atta, S., Kerubo, M. L., Sabisa, N. E., Adu-Gyamfi, D., Appiah, S., & Onyishi, E. (2025). Environmental corrosion and long-term degradation of crystalline silicon solar cells: Mechanisms, climate effects and mitigation strategies. Current Journal of Applied Science and Technology, 44(10), 9–18.
78) Atta, S., Onyishi, E. E., Appiah, S. A., & Adenuga, O. M. (2024). Theoretical analysis of charge transport and recombination losses in thin film solar cell architectures. International Journal of Engineering and Modern Technology, 10(10), 147–170. https://doi.org/10.56201/ijemt.v10.no10.2024.pg147.170
79) Atta, S., Tofah, P., & Adenuga, O. M. (2021). Advances in non-destructive testing methods for weld integrity evaluation in high-capacity power infrastructure. International Journal of Engineering and Modern Technology, 7(1), 20–47. https://doi.org/10.56201/ijemt.vol.7.no1.2021.pg20.47
80) Atta, S., Tofah, P., Kankam, M. A., & Adenuga, O. M. (2020). A critical review of NDT-based integrity management and corrosion risk assessment strategies for refinery process equipment. International Journal of Engineering and Modern Technology, 6(2), 19–46. https://doi.org/10.56201/ijemt.vol.6.no2.2020.pg19.46
81) Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2018). A systematic review of CI/CD pipeline strategies in Salesforce DevOps: Tools, practices, and deployment outcomes. Iconic Research and Engineering Journals, 2(6).
82) Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
83) Borah, S., Aliliele, K. C., Rakshit, S., & Vajjhala, N. R. (2022). Applications of artificial intelligence in software testing. In P. K. Mallick et al. (Eds.), Cognitive informatics and soft computing (Lecture Notes in Networks and Systems, Vol. 375, pp. 727–736). Springer. https://doi.org/10.1007/978-981-16-8763-1_60
84) Broder, A. Z. (1997). On the resemblance and containment of documents. Proceedings of the Compression and Complexity of Sequences, 21–29. https://doi.org/10.1109/SEQUEN.1997.666900
85) Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
86) Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., St. John, R., … Kurzweil, R. (2018). Universal sentence encoder. arXiv:1803.11175.
87) Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. Proceedings of the 34th ACM Symposium on Theory of Computing, 380–388.
88) Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer.
89) Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. Proceedings of the IJCAI Workshop on Information Integration on the Web, 73–78.
90) Dagodzo, D. (2018a). A conceptual framework for UAV integration into national power grid inspection programs. Iconic Research and Engineering Journals, 2(5), 391–412. https://doi.org/10.64388/IREV2I5-1716082
91) Dagodzo, D., Ahiaeke Patrick, M. C., & Aliliele, C. (2022). AI and deep learning for vegetation classification in power corridor management: A review. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 668–698. https://doi.org/10.32628/CSEIT2281227
92) Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423
93) Dosunmu, A. A., & Ogundele, P. O. (2021). Incident response and digital forensics strategies for rapid cyber attack containment. Gyanshauryam, International Scientific Refereed Research Journal, 4(4), 239–258.

94) Dosunmu, A. A., & Ogundele, P. O. (2022). Threat intelligence integration frameworks supporting proactive enterprise cybersecurity decision making. Gyanshauryam, International Scientific Refereed Research Journal, 5(3), 397–416.
95) Dosunmu, A. A., & Ogundele, P. O. (2023). Cyber threat actor analysis models for strategic enterprise security planning. Shodhshauryam, International Scientific Refereed Research Journal, 6(5), 513–531.
96) Dosunmu, A. A., & Ogundele, P. O. (2024a). Breach and attack simulation frameworks for continuous validation of enterprise security controls. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(3), 1100–1119.
97) Dosunmu, A. A., & Ogundele, P. O. (2024b). Cyber risk quantification models for prioritizing enterprise security investment decisions. International Journal of Multidisciplinary Research and Growth Evaluation, 5(6), 1777–1785. https://doi.org/10.54660/.IJMRGE.2024.5.6.1777-1785
98) Dosunmu, A. A., & Ogundele, P. O. (2024d). Threat informed defence engineering models for measuring security control effectiveness at scale. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2847–2858.
99) Dosunmu, A. A., & Ogundele, P. O. (2025a). Adversary simulation design frameworks for proactive cyber defense in complex environments. International Journal of Computer Science and Mathematical Theory, 11(12), 193–209. https://doi.org/10.56201/ijcsmt.vol.11.no12.2025.pg193.209
100) Dosunmu, A. A., & Ogundele, P. O. (2025c). Security orchestration and automation models for accelerating incident detection and response. Computer Science and IT Research Journal, 6(11), 878–894. https://doi.org/10.51594/csitrj.v6i11.2163
101) Dosunmu, A. A., & Ogundele, P. O. (2026a). Deception based defense architectures for disrupting advanced persistent threat operations. Engineering and Technology Journal, 11(1), 8475–8487. https://doi.org/10.47191/etj/v11i01.07
102) Eboh, E. E., & Aliliele, C. (2025b). Predictive analytics systems for investment risk monitoring in SME FinTech companies and cross-border capital markets. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(2), 3969–4010. https://doi.org/10.32628/CSEIT25113398
103) Edivri, J., & Oteri, O. (2022). Predictive capacity planning and resource utilization forecasting models for multi-program and multi-stakeholder IT portfolios. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(4), 826–845.
104) Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 1–16.
105) Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of KDD, 226–231.
106) Fawehinmi, Y. O., Okereke, G. M. T., Williams, P., Chijioke-Churuba, J. N., & Asare-Badu, B. (2026). AI-human collaboration in education and psychology: Personalised learning, language acquisition and student support systems. Journal of Political Science and International Relationship, 3(1), 54–64. https://doi.org/10.54536/jpsir.v3i1.6998
107) Fobellah, A. N. (2025a). Cognitive biases in financial decision-making: Implications for audit and risk management in large corporations: A conceptual review. International Journal of Science and Research Archive, 16(2), 17–22. https://doi.org/10.30574/ijsra.2025.16.2.2273
108) Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of EMNLP, 6894–6910.
109) Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. International Journal of Computer Applications, 68(13), 13–18.
110) Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. Wiley.
111) Hussain, N., & Adebayo, A. (2026). The role of artificial intelligence in strengthening modern cybersecurity systems. World Journal of Innovation and Modern Technology, 10(6), 125–181. https://doi.org/10.56201/wjimt.v10.no6.2026.pg125.181
112) Jimoh, H. O., & Ahmed, M. O. (2024). Analyzing Network Time Protocol (NTP) based amplification DDoS attack and its mitigation techniques. Journal of Digital Innovations and Contemporary Research in Science, Engineering and Technology, 12(2), 17–24. https://doi.org/10.22624/AIMS/DIGITAL/V11N2P2x
113) Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572
114) Komi, N. M. (2024). Turning Social License into A Measurable Variable: Predicting Community Trust in Renewable Energy Development. International Journal of Engineering and Modern Technology, 10(11), 239–296. https://doi.org/10.56201/ijemt.v10.no11.2024.pg239.296
115) Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. Data & Knowledge Engineering, 69(2), 197–210.
116) Lawal, O. A., & Oduleye, T. E. (2021a). A conceptual decision model for capital allocation using financial analytics. Gyanshauryam, International Scientific Refereed Research Journal, 4(2), 269–295.
117) Lawal, O. A., & Oduleye, T. E. (2021b). Aligning financial planning analytics with corporate strategy: A conceptual integration model. Shodhshauryam, International Scientific Refereed Research Journal, 4(3), 319–346.
118) Lawal, O. A., & Oduleye, T. E. (2023a). A review of decision analytics models for sustainable profitability in technology firms. Gyanshauryam, International Scientific Refereed Research Journal, 6(6), 455–475.
119) Lawal, O. A., & Oduleye, T. E. (2023b). Behavioral financial analytics: A conceptual model for explaining enterprise performance. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2590–2604.
120) Lawal, O. A., & Oduleye, T. E. (2026). Integrated financial intelligence architectures: A conceptual model for global scalability. Gulf Journal of Advance Business Research, 4(1), 10–33. https://doi.org/10.51594/gjabr.v4i1.197
121) Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), 707–710.

122) Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824–836.
123) Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.
124) McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 205.
125) Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.
126) Ogunwola, T. A. (2026). Security and resilience considerations for software-defined wide area network deployments in multi-site enterprise environments. International Advanced Research Journal in Science, Engineering and Technology, 12(1). https://doi.org/10.17148/iarjset.2025.12150
127) Ogunwola, T. A., & Alozie, C. (2026a). Enterprise firewall modernization for business continuity: A framework for secure migration and operational stability. Computer Science & IT Research Journal, 7(5), 347–361. https://doi.org/10.51594/csitrj.v7i5.2287
128) Ogunwola, T. A., & Deborah, F. O. (2026). Enterprise network resilience as a national security imperative: Strategies for protecting United States critical infrastructure. International Advanced Research Journal in Science, Engineering and Technology, 12(10). https://doi.org/10.17148/iarjset.2025.121048
129) Ogunwola, T. A., Owoyemi, T. A., & Iyanda, C. (2026). Cybersecurity and network integration risk management during corporate acquisitions and infrastructure consolidation. Computer Science & IT Research Journal, 7(6), 389–404. https://doi.org/10.51594/csitrj.v7i6.2305
130) Onwubuya, L. I., Fawehinmi, Y. O., Sunday, D., Igwenagu, M. O., Adeniran, A. O., & Agyapong, E. A. (2026). Adaptive learning systems powered by artificial intelligence for STEM education improvement. Asian Journal of Education and Social Studies, 52(6), 553–566. https://doi.org/10.9734/ajess/2026/v52i63115
131) Orise, C., & Niniola, S. (2023). Privacy, safeguarding, and ethical data practices in healthcare EdTech: A conceptual model for compliance and trust. Shodhshauryam, International Scientific Refereed Research Journal, 6(1), 507–529.
132) Orise, C., & Niniola, S. (2025). Data governance in digital health learning systems: A systematic review of implementation frameworks and best practices. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(5), 520–538.
133) Orise, C., & Niniola, S. (2026). From classroom to clinic: Digital transformation in health professional education and training delivery. International Journal of Health and Pharmaceutical Research, 11(6), 52–64. https://doi.org/10.56201/ijhpr.vol.11.no6.2026.pg52.64
134) Oyeleke, A. V., Eze, F. C., & Asiedu, W. (2026). Reliability-centered maintenance strategies for minimizing downtime and maximizing performance in high-density GPU cluster environments. International Journal of Engineering Technology Research & Management, 10(7), 18–38.
135) Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of EMNLP, 1532–1543. https://doi.org/10.3115/v1/D14-1162
136) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2024). Modeling the impact of impedance mismatch on wireless link performance: A framework for prototype development and system optimization. International Journal of Computer Science and Mathematical Theory, 10(2), 62–116. https://doi.org/10.56201/ijcsmt.v10.no2.2024.pg62.116
137) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2025a). Predicting throughput degradation in Wi-Fi 6 networks under varying channel conditions: A physical layer performance model. International Journal of Engineering and Modern Technology, 11(12), 205–264. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.264
138) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2025b). Test automation in wireless hardware engineering: A comprehensive review of scripting frameworks and instrument control strategies. International Journal of Engineering and Modern Technology, 11(10), 405–464. https://doi.org/10.56201/ijemt.vol.11.no10.2025.pg405.464
139) Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP, 3982–3992. https://doi.org/10.18653/v1/D19-1410
140) Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.
141) Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. Journal of Big Data, 7(1), 41.
142) Sauerwein, C., Sillaber, C., Mussmann, A., & Breu, R. (2017). Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives. Proceedings of the International Conference on Wirtschaftsinformatik, 837–851.
143) Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423.
144) Skopik, F., Settanni, G., & Fiedler, R. (2016). A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing. Computers & Security, 60, 154–176.
145) Steorts, R. C., Hall, R., & Fienberg, S. E. (2016). A Bayesian approach to graphical record linkage and de-duplication. Journal of the American Statistical Association, 111(516), 1660–1672.
146) Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., & Zhang, J. (2023). Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. IEEE Communications Surveys & Tutorials, 25(3), 1748–1774.
147) Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & Security, 72, 212–233. https://doi.org/10.1016/j.cose.2017.09.001
148) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

149) Wagner, C., Dulaunoy, A., Wagener, G., & Iklody, A. (2016). MISP: The design and implementation of a collaborative threat intelligence sharing platform. Proceedings of the ACM Workshop on Information Sharing and Collaborative Security, 49–56.
150) Wagner, T. D., Mahbub, K., Palomar, E., & Abdallah, A. E. (2019). Cyber threat intelligence sharing: Survey and research directions. Computers & Security, 87, 101589. https://doi.org/10.1016/j.cose.2019.101589

Corresponding Author: Abolaji Adebayo