BDAA 2025 Innsbruck

Big Data Analytics & Applications

Proceedings of the 1st International Conference on Big Data Analytics & Applications (BDAA' 2025)

25-27 November 2025
Innsbruck, Austria

IFSA

Big Data Analytics & Applications

Proceedings of the First International Conference on Big Data Analytics & Applications (BDAA' 2025)

25-27 November 2025

Innsbruck, Austria

Edited by Sergey Y. Yurish

IFSA Publishing

Sergey Y. Yurish, *Editor*
BDAA' 2025 Conference Proceedings

E-mail (for orders and customer service enquires): ifsa.books@sensorsportal.com

Visit our Home Page on http://www.sensorsportal.com

ISBN: 978-84-09-78845-3
BN-20251120-XX
BIC: GPH

Contents

Foreword

It is my great pleasure to present the Proceedings of the First International Conference on Big Data Analytics & Applications (BDAA' 2025), held on 25–27 November 2025 in Innsbruck, Austria. This inaugural edition marks an important milestone for our community, bringing together researchers, engineers, practitioners, and innovators working across the rapidly expanding domains of big data engineering, intelligent analytics, machine learning, data privacy, high-performance systems, and domain-focused data applications.

The extraordinary growth of data-driven technologies continues to reshape science, industry, government, and society. Yet this growth also brings new responsibilities—ensuring trust, security, transparency, and efficiency in modern data ecosystems. BDAA' 2025 was established to provide a focused, high-quality platform for addressing these emerging challenges. Our aim is to bridge foundational advances in algorithms, models, and architectures with real-world applications in fields such as health, cybersecurity, education, social platforms, and large-scale scientific projects.

The papers included in these proceedings reflect the diverse and multidisciplinary character of the conference. They cover topics such as knowledge-graph-enhanced academic resource discovery, privacy-preserving federated learning, face de-identification and generative pseudonymization, secure microservice-based authorization frameworks, and advanced analytical methods for cloud environments, education, and collaborative research infrastructures. Despite their variety, all contributions share a common ambition: to advance the state of the art in extracting meaningful, reliable, and actionable insights from complex data. Each paper has undergone careful peer review to ensure scientific quality and relevance.

I would like to express my sincere appreciation to all authors for their high-quality submissions, and to the members of the Program Committee and external reviewers for their dedication and rigorous evaluations. My thanks also go to our invited speakers and session chairs, whose expertise enriched the scientific discussions, and to all participants whose engagement made BDAA' 2025 a vibrant and productive event.

Finally, I extend my gratitude to IFSA Publishing for their continued support and for preparing this volume, which I hope will serve as a useful resource for researchers and practitioners worldwide. I am confident that the ideas presented here will inspire further innovations and collaborations, and will contribute to shaping the future of big data analytics and its transformative applications.

Prof., Dr. Sergey Y. Yurish
BDAA' 2025 Chairman

(006)

AcademicRAG: A Knowledge Graph-enhanced Framework for Intelligent Academic Resource Discovery

Zhuchenyang Liu [1], Shuhua Chen [2], Shiva Sander Tavallaey [3,4], Fredrik Heintz [5] and Elias Zea [4]
[1] Aalto University, School of Electrical Engineering, Espoo, Finland
[2] Alibaba Cloud Computing Co., Ltd., Hangzhou, China
[3] ABB AB Corporate Research, Västerås, Sweden
[4] KTH Royal Institute of Technology, Department of Engineering Mechanics, Stockholm, Sweden
[5] Linköping University, Department of Computer and Information Science, Linköping, Sweden
Tel.: +46 079 353 5523
E-mail: zhuchenyang.liu@aalto.fi

Summary: Traditional academic retrieval systems struggle to capture complex semantic relationships; this often leads to incomplete results and missed interdisciplinary connections. We present AcademicRAG [1], a novel knowledge graph-enhanced framework that addresses these limitations through two key innovations: clue-guided keyword generation and subgraph-based retrieval. Unlike GraphRAG's expensive community structures and LightRAG's limited one-hop neighbor retrieval, AcademicRAG builds complete subgraphs while anchoring keywords in actual graph content to prevent semantic drift. Additionally, it eliminates community regeneration overhead through clue-guided keyword indexing, enabling efficient incremental updates. Evaluations across agriculture and computer science domains demonstrate that our approach consistently outperforms three state-of-the-art baselines. We demonstrate AcademicRAG's practical versatility by powering a research literature assistant that leverages clue-guided querying and subgraph navigation for precise, context-aware paper discovery, and a course discovery system that uses dynamic subgraph analysis to generate personalized learning pathways and optimize course curricula in the acoustics domain.

Keywords: Knowledge graph, Retrieval-augmented generation, Graph-based RAG, Academic resource discovery, Subgraph retrieval, Literature review, Course discovery.

1. Introduction

The exponential growth of digital academic content has transformed scholarly communication, with academic databases now containing millions of papers across diverse disciplines [1]. However, this unprecedented scale creates significant challenges in knowledge discovery and synthesis, particularly in academic contexts where understanding depends on grasping complex relationships between entities, methodologies, and concepts [2].

Traditional retrieval systems demonstrate fundamental limitations in academic knowledge discovery. Sparse retrieval methods like BM25, while computationally efficient, struggle with terminology variations and fail to capture semantic relationships inherent in academic knowledge [3]. Dense retrieval methods using Transformer-based encoders improve semantic understanding but cannot preserve the intricate relational structures that define academic knowledge hierarchies, prerequisite chains, and interdisciplinary connections [4].

Recent Retrieval-Augmented Generation (RAG) approaches, though promising, inherit these structural limitations. Conventional RAG treats academic documents as isolated entities with flat vector representations, failing to model the complex relationship networks essential for comprehensive academic understanding [5]. Four main RAG paradigms have emerged – Sequential, Branching, Conditional, and Loop RAG [6] – but none adequately addresses the spectrum of academic information needs.

Graph-based RAG architectures attempt to address these limitations but face critical trade-offs [7]. GraphRAG [8] employs hierarchical community structures, achieving improvement in critical dimensions over traditional RAG, but requires high computational complexity for community generation, making incremental updates prohibitively expensive. LightRAG [9] reduces computational overhead through dual-level retrieval but limits relationship discovery to one-hop neighbors, fundamentally constraining multi-hop dependency capture crucial for academic synthesis. Other frameworks (GRAG [10], HybridRAG [11], CG-RAG [12]) face similar challenges: high computational requirements, limited multi-hop inference capabilities, and manual knowledge graph curation needs.

We present AcademicRAG, a knowledge graph-enhanced RAG framework specifically designed for intelligent academic resource discovery. Our approach introduces two key technical innovations that address the fundamental limitations of existing graph-based RAG frameworks. Firstly, we develop clue-guided keyword generation that anchors keyword extraction processes in actual graph content, thereby

[1] Code and additional implementation details are available at: https://github.com/shua-chen/academicRAG

preventing semantic drift and hallucination issues that commonly plague generative retrieval systems when operating without grounded constraints. Secondly, we implement subgraph-based retrieval mechanisms that construct complete relationship networks efficiently, eliminating the need for computationally expensive community structures while avoiding the restrictive single-hop neighbor limitations that constrain existing approaches.

Our contributions are threefold. First, we develop a dual-flow architecture transforming unstructured academic content into quarriable knowledge graphs through iterative entity-relationship extraction. Second, we introduce efficient algorithms for local subgraph construction and global relationship discovery that capture multi-hop dependencies while enabling incremental updates, validated through substantive evaluation. Third, we demonstrate practical versatility through two real-world applications addressing distinct academic information needs across different user groups.

Comprehensive technical details, implementation specifics, and extended case studies are available in the complete documentation [13].

2. AcademicRAG Framework

AcademicRAG operates through two primary flows: Index Flow for knowledge graph construction and Query Flow for information retrieval, as illustrated in Fig. 1. This dual-flow architecture transforms unstructured academic content into structured knowledge representations while enabling sophisticated query processing that captures both local detail and global context.

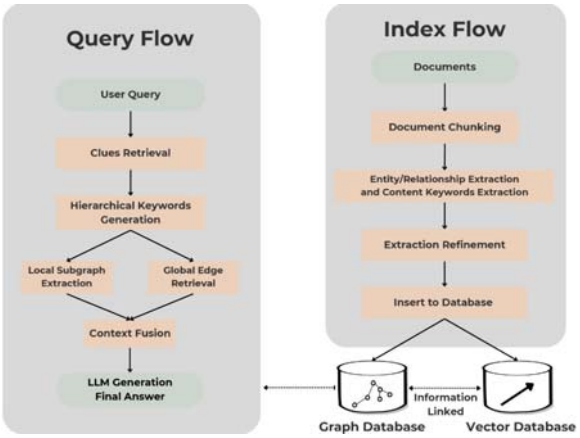


Fig. 1. AcademicRAG Framework Flowchart.

2.1. Index Flow

The Index Flow processes academic documents through a comprehensive multi-stage pipeline designed to preserve semantic relationships while enabling efficient retrieval. The process begins with document ingestion and structure-aware chunking that respects document organization, particularly important for academic content with inherent hierarchical structure such as research papers and course syllabi.

Each document chunk undergoes entity and relationship extraction using domain-adapted LLMs for academic content. The extraction process simultaneously identifies three critical components:

$$(E_i, R_i, K_i) = \mathrm{LLM}(c_i), \tag{1}$$

where c_i represents document chunks, E_i and R_i are extracted entities and relationships, and K_i are content keywords capturing global themes.

Next, the framework employs a multi-pass iterative refinement mechanism that maximizes entity and relationship extraction accuracy. Unlike single-pass extraction methods, this approach performs systematic re-extraction to capture complex academic relationships that may be missed in initial processing.

The extracted information is stored in a dual-database architecture: the Graph Database preserving structural relationships and the Vector Database storing semantic embeddings. This hybrid approach enables both structured traversal and similarity-based retrieval, providing comprehensive coverage for complex academic queries.

2.2. Query Flow

Upon receiving a user query Q, the system implements a sophisticated multi-stage retrieval strategy that addresses semantic drift issues common in knowledge graph querying. The process begins with clue-guided keyword generation, which anchors keyword extraction in actual database content to prevent hallucination:

$$K^{\text{content}} = \{\, k \in K \mid \mathrm{sim}(Q, k) \geq \tau \,\}, \tag{2}$$

where K is the set of indexed content keywords, Q is the user query, and τ is a similarity threshold. The function $\mathrm{sim}(Q, k)$ computes the embedding cosine similarity between query and keyword. These clues K^{content}, anchored in actual database content, guide the LLM to generate hierarchical keywords that avoid semantic drift by combining semantic information from the user query with natural language instructions:

$$(K_h, K_l) = \mathrm{LLM}(Q, K^{\text{content}}),. \tag{3}$$

where $K_h$represents high-level conceptual keywords and K_l represents low-level contextual keywords. Low-level keywords drive local subgraph extraction through the structured process detailed in Fig. 2. Unlike LightRAG's [9] one-hop limitation, our approach constructs complete subgraphs that capture multi-hop relationships between specific entities, enhancing retrieval depth and knowledge integration.

High-level keywords enable global relationship discovery through the systematic process illustrated in

Fig. 3. The framework queries the vector database to identify edges correlating with broader conceptual frameworks, eliminating GraphRAG's [8] community structure overhead while maintaining comprehensive global knowledge discovery.

Algorithm 1 Local Information Extraction based on Subgraph

Require: Low-level keywords $\mathcal{K}_l$, Knowledge graph $\mathcal{G} = (V, E)$
Ensure: Local entities E_{local}, relationships R_{local}, text units T_{local}

Node Matching: Match graph nodes against $\mathcal{K}_l$ using vector similarity metrics
$V_{\text{matched}} \leftarrow \{v \in V \mid \max_{k \in \mathcal{K}_l} \text{sim}(v, k) \geq \theta_{\text{node}}\}$
Subgraph Construction: Generate subgraph by finding shortest paths between matched nodes
$\mathcal{G}_{\text{sub}} \leftarrow \text{ShortestPaths}(V_{\text{matched}}, \mathcal{G})$
Pruning: Filter edges based on semantic relevance thresholds
$E_{\text{filtered}} \leftarrow \{e \in E_{\text{sub}} \mid \max_{k \in \mathcal{K}_l} \text{sim}(e.\text{description}, k) \geq \theta_{\text{edge}}\}$
Remove isolated nodes after edge filtering
Text Unit Extraction: Extract and rank textual chunks from retained edges
$T_{\text{chunks}} \leftarrow \text{ExtractTextUnits}(E_{\text{filtered}})$
Rank T_{chunks} by frequency within subgraph
Edge Expansion: Iteratively expand subgraph with one-hop neighbors
$E_{\text{expanded}} \leftarrow E_{\text{filtered}} \cup \text{OneHopNeighbors}(V_{\text{matched}}, \theta_{\text{relevance}})$
Data Truncation: Truncate all data to meet computational constraints
while $\text{TokenCount}(E_{\text{local}}, R_{\text{local}}, T_{\text{local}}) > \tau_{\max}$ **do**
  Remove lowest-ranked elements
end while
return $(E_{\text{local}}, R_{\text{local}}, T_{\text{local}})$

Fig. 2. Local Information Extraction based on Subgraph.

Algorithm 2 Global Information Extraction

Require: High-level keywords $\mathcal{K}_h$, Knowledge graph $\mathcal{G} = (V, E)$, Vector database $\mathcal{D}_{\text{vector}}$
Ensure: Global entities E_{global}, relationships R_{global}, text units T_{global}

Edge Retrieval: Query vector database for edges correlating with $\mathcal{K}_h$
$E_{\text{candidate}} \leftarrow \{e \in E \mid \max_{k \in \mathcal{K}_h} \text{sim}(\text{embed}(e), \text{embed}(k)) \geq \theta_{\text{global}}\}$
Prioritize edges encapsulating conceptual relationships
Node Retrieval: Extract nodes related to retrieved edges
$V_{\text{related}} \leftarrow \{v \in V \mid \exists e \in E_{\text{candidate}}, v \in \{e.\text{source}, e.\text{target}\}\}$
Sort V_{related} by node degree
Edge Data Processing: Analyze edge metadata
for $e \in E_{\text{candidate}}$ **do**
  $e.\text{degree} \leftarrow \text{degree}(e.\text{source}) + \text{degree}(e.\text{target})$
  $e.\text{weight} \leftarrow \text{SemanticConnectionStrength}(e)$
end for
Integration and Ranking: Rank edges using composite metric
$\text{score}(e) \leftarrow \alpha \cdot \text{RelevanceScore}(e, \mathcal{K}_h) + \beta \cdot e.\text{weight}$
$E_{\text{ranked}} \leftarrow \text{Sort}(E_{\text{candidate}}, \text{score})$
Text Unit Extraction: Extract text chunks associated with edges
$T_{\text{chunks}} \leftarrow \text{ExtractTextUnits}(E_{\text{ranked}})$
Rank by similarity between edges and $\mathcal{K}_h$
Data Truncation: Truncate to meet computational constraints
while $\text{TokenCount}(E_{\text{global}}, R_{\text{global}}, T_{\text{global}}) > \tau_{\max}$ **do**
  Remove lowest-ranked elements
end while
return $(E_{\text{global}}, R_{\text{global}}, T_{\text{global}})$

Fig. 3. Global Information Extraction.

The final synthesis stage combines retrieved information through sophisticated contextual fusion to generate comprehensive responses. The system integrates locally extracted subgraph information with globally retrieved relationship networks, merging entities, relationships, and supporting textual evidence into a unified contextual framework. This integrated information is then embedded within a domain-specific prompt template designed for academic content processing. The large language model processes this enriched context to generate the final answer, leveraging both the granular detail from local subgraphs and the broader conceptual understanding from global edge networks. This fusion approach ensures that responses maintain semantic coherence while incorporating multiple levels of contextual depth, from specific entity relationships to overarching domain knowledge.

3. Evaluation and Results

3.1. Experiment Settings

We evaluated AcademicRAG using a subset of the UltraDomain benchmark [14], selecting agriculture and computer science domains for their distinct knowledge structures and citation patterns. Following Microsoft's GraphRAG methodology [8], we generated five synthetic academic personas per domain (e.g., "PhD Student in NLP," "Organic Farming Consultant"), each assigned 25 diverse queries reflecting common academic information needs, resulting in 125 evaluation queries per domain.

We compared AcademicRAG against three representative baselines: NaiveRAG [5], GraphRAG [8] and LightRAG [9]. These baselines were selected to provide comprehensive coverage of current RAG paradigms, from flat document processing to sophisticated graph-based approaches.

All frameworks used identical configurations with Qwen2.5-72B [15] for both knowledge graph construction and response generation, ensuring fair comparison. Evaluation employed DeepSeek-R1 [16] as an LLM judge conducting pairwise comparisons across four dimensions:

- **Comprehensiveness**: Coverage of relevant information and depth of analysis;
- **Diversity**: Variety of perspectives and information sources;
- **Empowerment**: Usefulness for decision-making and further research;
- **Overall**: Holistic assessment integrating all dimensions for comprehensive performance.

To mitigate positional bias, we alternated answer placement across three trials and computed averaged win rates. While LLM-as-Judge represents current best practice for high-level semantic evaluation, we acknowledge the need for human expert validation in future work.

3.2. Evaluation Results

As shown in Table 1, AcademicRAG demonstrated consistent superiority across both domains and all evaluation dimensions. Our framework achieved overall win rates of 57.2 %, 54.8 %, and 52.4 % against NaiveRAG, LightRAG, and GraphRAG respectively in agriculture, with even more pronounced advantages in computer science (77.5 %, 56.6 %, and 53.6 %). The significant performance disparity between domains suggests that field-specific characteristics – including terminology standardization and conceptual hierarchy – influence framework effectiveness, with computer science's more structured knowledge representation facilitating superior retrieval quality.

The results validate the core technical innovations of our dual-level retrieval approach. AcademicRAG's subgraph-based retrieval enables extraction of richer, more structured information compared to flat document approaches, while clue-guided keyword generation ensures semantic alignment with indexed content, preventing hallucination issues in keyword generation. Notably, while consistently outperforming LightRAG across all dimensions, the comparison with GraphRAG reveals that our framework excels in comprehensiveness and empowerment – the most critical metrics for academic information discovery – demonstrating that complete subgraph construction captures essential multi-hop relationships that community-based approaches may overlook.

3.3. Ablation Study

To validate our technical innovations, we conducted ablation studies examining the individual contributions of clue-guided keyword generation and subgraph-based retrieval. Table 2 presents the results of removing each component against GraphRAG as baseline. Removing clues resulted in mixed performance changes, with diversity scores declining from 50.8 % to 48.4 % in agriculture and from 46.8 % to 43.0 % in computer science, while surprisingly showing improvements in empowerment (54.0 % to 60.8 % in agriculture). The subgraph removal variant demonstrated similar complexity, with reduced comprehensiveness in computer science (51.2 % to 49.2 %) but improved overall performance in the same domain (53.6 % to 55.2 %).

These mixed ablation results reflect the inherent complexity of graph-based RAG evaluation and the nuanced trade-offs between our technical components. The unexpected improvements in certain configurations can be attributed to several factors. Firstly, the subjective nature of LLM-as-Judge evaluation may lead to scenarios where simpler responses are occasionally preferred for specific query types, potentially masking the true effectiveness of more sophisticated retrieval mechanisms. Secondly, the interaction effects between components create non-linear performance relationships, wherein the removal of one component may inadvertently optimize the remaining system configuration for particular evaluation criteria. Thirdly, GraphRAG's community-based approach introduces variable noise levels that differentially affect comparisons, as the quality and coherence of generated communities can vary significantly across different knowledge domains and corpus structures.

4. Applications

To demonstrate AcademicRAG's practical versatility beyond the evaluated domains, we developed two real-world applications that address distinct academic information needs across different user groups while validating the framework's adaptability to diverse academic contexts.

Table 1. Comparative Win Rates (%) of AcademicRAG against Baseline Models Across Two Domains.

Framework	Agriculture				Computer Science			
	Comp	Div	Emp	Overall	Comp	Div	Emp	Overall
NaiveRAG	62.8 %	58.0 %	51.6 %	**57.2 %**	67.1 %	83.2 %	75.7 %	**77.5 %**
LightRAG	50.0 %	55.2 %	58.4 %	**54.8 %**	56.2 %	57.8 %	59.8 %	**56.6 %**
GraphRAG	52.4 %	50.8 %	54.0 %	**52.4 %**	51.2 %	46.8 %	55.2 %	**53.6 %**

Note: Comp = Comprehensiveness, Div = Diversity, Emp = Empowerment. Win rates were computed based on pairwise comparisons conducted by a Large Language Model (LLM-as-Judge) over 125 domain-specific queries, following common practices in graph-based RAG framework evaluation [8, 9].

Table 2. Ablation Studies on Win Rates (%) Compared with GraphRAG.

Settings	Agriculture				Computer Science			
	Comp	Div	Emp	Overall	Comp	Div	Emp	Overall
AcademicRAG	52.4 %	50.8 %	54.0 %	52.4 %	51.2 %	46.8 %	55.2 %	53.6 %
-clues	53.6 %	48.4 %	60.8 %	57.2 %	51.0 %	43.0 %	54.6 %	51.8 %
-subgraph	50.8 %	51.6 %	52.8 %	53.6 %	49.2 %	48.0 %	59.3 %	55.2 %

Note: Comp = Comprehensiveness, Div = Diversity, Emp = Empowerment. "-clues" and "-subgraph" represent ablated versions removing the respective retrieval mechanisms, with "-subgraph" using one-hop retrieval only.

4.1. Course Discovery System

We adapted AcademicRAG for educational content navigation using 29 acoustics-related course syllabi from KTH Royal Institute of Technology. The system required minimal framework modifications, including education-specific entity schemas for courses and prerequisites, course-structure-preserving chunking strategies, and domain-tailored prompts optimized for educational content processing.

The resulting knowledge graph comprised 1211 nodes and 2252 edges with strong local clustering, demonstrating effective capture of course relationships and prerequisite structures. The largest

connected component encompassed 94.63 % of all nodes, indicating well-integrated knowledge representation across the curriculum.

Case study evaluation focused on two distinct user groups with different information needs. For students, the system demonstrated sophisticated curriculum mapping capabilities by generating logical learning progressions that respect educational hierarchies, providing precise prerequisite identification for advanced courses, and delivering multi-dimensional course recommendations balancing theoretical foundations with practical laboratory components. For faculty and administrators, the system successfully classified all courses by academic level and offering term, identified content overlaps through learning outcome analysis, and generated specific integration proposals with clear implementation methodologies. These capabilities enable evidence-based curriculum optimization and resource allocation decisions that transform abstract analyses into executable administrative workflows.

4.2. Research Literature Assistant

The Research Literature Assistant was developed using over 80 research papers across multiple academic domains, with framework adaptations including specialized prompts for academic content, automated title and introduction extraction methods, and structure-aware chunking strategies that preserve paper organization.

We conducted comprehensive evaluation through literature review scenario testing, focusing on blind face restoration within computer vision as a representative domain with methodological diversity. The assistant demonstrated exceptional performance across multiple dimensions: identifying major research trends and specific gaps in the field, providing contextual paper recommendations with multi-faceted contribution analysis, extracting detailed technical information typically requiring manual paper review, and organizing literature into coherent methodological categories spanning fundamental works, domain-specific applications, and hybrid approaches.

User testing with graduate students revealed significant advantages over online LLM-based systems, particularly in accommodating larger paper corpora while maintaining higher factual accuracy. The system's grounding in indexed content minimized probability of hallucination issues common in general-purpose LLMs, ensuring responses remained anchored to actual paper content. Key strengths included effective research gap identification, precise semantic search capabilities across multiple papers, and comprehensive paper relationship mapping that revealed hidden connections between publications.

4.3. Practical Insights and Limitations

Both applications demonstrated the framework's cross-domain adaptability through its modular design and LLM-driven entity-relationship extraction with domain-specific prompt engineering. However, several practical limitations emerged that warrant acknowledgment. Firstly, the framework exhibits weaker comprehension of non-textual elements such as figures, mathematical formulas, and complex diagrams, which constitute critical components of academic content across many disciplines. Secondly, the system demonstrates significant dependency on underlying LLM capabilities, with quality variations in the base model directly impacting system reliability and output consistency. Thirdly, our experimental validation was constrained by computational resource limitations, restricting the scale of evaluation to relatively small-scale deployments, which may limit the generalizability of findings to large-scale institutional applications. Despite these constraints, both applications provided substantial value for academic workflows, particularly in educational pathway planning and literature synthesis tasks, validating AcademicRAG's potential for transforming academic resource discovery across diverse institutional contexts while highlighting areas requiring future research attention.

5. Conclusions

This work presents AcademicRAG, a novel knowledge graph-enhanced RAG framework that significantly advances academic resource discovery through two key technical innovations: clue-guided keyword generation and subgraph-based retrieval. By anchoring keyword extraction in actual graph content and constructing complete subgraphs, our framework addresses fundamental limitations in existing approaches while achieving superior computational efficiency compared to GraphRAG's community-based architecture. Comprehensive evaluation across agriculture and computer science domains demonstrates consistent performance advantages over state-of-the-art baselines, with win rates exceeding 50 % across all evaluation dimensions. The framework's dual-level retrieval approach effectively balances local detail extraction with global context discovery, particularly excelling in comprehensiveness and user empowerment – the most critical metrics for academic information discovery.

The practical utility of AcademicRAG is validated through successful deployment of two real-world applications: a Course Discovery System enabling sophisticated curriculum analysis, and a Research Literature Assistant providing factually grounded literature synthesis capabilities. Both applications demonstrate the framework's versatility and cross-domain adaptability through minimal modifications, successfully addressing distinct information needs for students, faculty, and researchers. The applications revealed important insights about academic workflows while highlighting the framework's ability to transform unstructured academic content into semantically rich, queryable

knowledge representations that support evidence-based decision-making.

Despite these advances, several limitations warrant acknowledgment, including significant dependency on underlying LLM capabilities, weaker comprehension of non-textual elements, and restricted experimental validation scale. Future work should prioritize multi-modal content processing, scalability optimization, and comprehensive human expert evaluation to strengthen the foundation for next-generation academic resource discovery systems.

Acknowledgements

The computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. We specifically acknowledge the computational resources allocated through NAISS project "GraphRAG for Academic Resource Discovery: A Dual-Application Framework", which provided essential GPU computing power on the Alvis system at C3SE and storage resources on Mimer, enabling the extensive experimentation and evaluation presented in this work.

Elias Zea gratefully acknowledges the financial support of the Swedish Research Council, Grant No. 2020-04668. This work was also partially supported by grants from the Excellence Center at Linköping-Lund for Information Technology (ELLIIT), which we gratefully acknowledge.

References

[1]. W. Ammar, et al., Construction of the literature graph in semantic scholar, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 3 (Industry Papers), 2018, pp. 84-91.

[2]. S. Ji, S. Pan, E. Cambria, P. Marttinen, et al., A survey on knowledge graphs: representation, acquisition, and applications, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, Issue 2, 2022, pp. 494-514.

[3]. S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends in Information Retrieval*, Vol. 3, Issue 4, 2009, pp. 333-389.

[4]. J. Wang, et al., Utilizing BERT for information retrieval: survey, applications, resources, and challenges, *ACM Computing Surveys*, Vol. 56, Issue 7, 2024, pp. 1-33.

[5]. P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems*, Vol. 33, 2020, pp. 9459-9474.

[6]. X. Li, et al., From matching to generation: a survey on generative information retrieval, *ACM Transactions on Information Systems*, Vol. 43, Issue 3, 2025, 83.

[7]. B. Peng, et al., Graph retrieval-augmented generation: a survey*, arXiv preprint*, 2024, arXiv:2408.08921.

[8]. D. Edge, et al., From local to global: a graph RAG approach to query-focused summarization, *arXiv preprint*, 2024, arXiv:2404.16130.

[9]. Z. Guo, L. Xia, Y. Yu, T. Ao, et al., LightRAG: simple and fast retrieval-augmented generation, *arXiv preprint*, 2024, arXiv:2410.05779.

[10]. Y. Hu, Z. Lei, Z. Zhang, B. Pan, et al., GRAG: graph retrieval-augmented generation, *arXiv preprint*, 2024, arXiv:2405.16506.

[11]. B. Sarmah, et al., HybridRAG: integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction, in *Proceedings of the 5th ACM International Conference on AI in Finance*, 2024, pp. 608-616.

[12]. Y. Hu, Z. Lei, Z. Dai, A. Zhang, et al., CG-RAG: research question answering by citation graph retrieval-augmented LLMs, *arXiv preprint*, 2025, arXiv:2501.15067.

[13]. S. Chen, Z. Liu, AcademicRAG: knowledge graph enhanced retrieval-augmented generation for academic resource discovery, Master's Thesis, *KTH Royal Institute of Technology & Aalto University*, 2025.

[14]. H. Qian, P. Zhang, Z. Liu, K. Mao, et al., MemoRAG: moving towards next-gen RAG via memory-inspired knowledge discovery, *arXiv preprint*, 2024, arXiv:2409.05591.

[15]. A. Yang, et al., Qwen2.5 technical report, *arXiv preprint*, 2024, arXiv:2412.15115.

[16]. DeepSeek-AI, DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning, *arXiv preprint*, 2025, arXiv:2501.12948.

(007)

A Holistic Privacy-preserving Framework for the Federated Learning Lifecycle

Lucía Muñoz-Solanas, Amaia Gil-Lerchundi
Vicomtech, Basque Research and Technology Alliance (BRTA), San Sebastian, Spain
E-mail: lmunoz@vicomtech.org, agil@vicomtech.org

Summary: Federated Learning is vulnerable to privacy attacks on federated model updates. This paper introduces ASTER, a holistic, framework providing end-to-end privacy for the Federated Learning lifecycle. The architecture combines three cryptographic stages: an e-voting inspired mix-net for anonymous training submissions; Homomorphic Encryption for secure server-side inference; and Secret Sharing Schemes for threshold-based control over model retraining. This system prevents user-contribution linking while guaranteeing data confidentiality and model integrity, creating a robust framework for sensitive applications.

Keywords: Privacy-preserving, Federated learning, E-voting, Homomorphic encryption, Secret sharing.

1. Introduction

Federated Learning (FL) enhances privacy by training on decentralized data kept on user devices. However, this process remains vulnerable to privacy breaches, particularly from the central server. Such server, while correctly following the protocol, can still analyze received model parameters to infer sensitive information and link contributions back to specific users [1]. This threat is critical in sensitive domains, and this work will focus specifically on the energy sector, where building consumption patterns can reveal private user behaviors.

A key limitation of many existing solutions is that they focus on protecting a single phase of the FL process rather than providing end-to-end protection. This leaves critical vulnerabilities at different stages; during training, the link ability between a user and their update is a major risk; during inference, a client's private query data is exposed to the server; and throughout the model's lifecycle, a malicious actor or a small colluding group could unilaterally alter the model, compromising its integrity. While techniques like Differential Privacy (DP) [2] can obscure training data, they often harm model accuracy, and while Homomorphic Encryption (HE) can protect inference queries, it often introduces prohibitive computational overhead for the entire training process [3].

To address these multi-faceted threats in a unified manner, this paper proposes a holistic, multi-stage privacy-preserving framework, developed within the ASTER project, that safeguards sensitive data across its entire lifecycle: training, inference, and aggregation. The main contributions of this work are threefold, with each component designed to counter a specific threat:

- **To prevent user-contribution linking during training phase**, an anonymization scheme inspired by electronic voting protocol is introduced. Using a mix-net architecture with a hybrid ElGamal+AES encryption scheme, the central server can aggregate model parameters without linking them to the originating clients;
- **To protect client data during the inference phase**, HE is leveraged, enabling clients to obtain predictions from the global model using their private data without ever revealing it in plaintext to the server;
- **To secure the retraining process against malicious or unilateral alterations**, Secret Sharing Schemes (SSS) are employed. This mechanism ensures that the federated model can only be updated if a minimum threshold of participants collaboratively agrees, establishing a democratic and robust control over the model.

2. Related Work

While the federated paradigm inherently improves privacy, it remains vulnerable to attacks such as membership inference [4] and model inversion [5]. To mitigate these risks, several Privacy-Enhancing Techniques (PETs) have been proposed, though they often focus on securing a single phase of the FL process.

A foundational approach is secure aggregation, where the server computes the sum of client updates without inspecting individual contributions. This is often implemented with SSS [6]. While this effectively protects the content of the updates, it does not anonymize the participants. The server still knows which clients are participating, which can be a source of information leakage. ASTER addresses this gap directly with its e-voting inspired mix-net, which anonymizes the source of contributions before they reach the server.

Another line of work explores HE. The application of HE to protect the inference phase in FL is a well-established research direction, with multiple studies exploring its use to allow clients to receive predictions on their private data without revealing it to

the server [7]. However, it is widely recognized that the primary limitation of HE is its significant computational overhead, which can make it impractical for latency-sensitive applications [8]. While these works provide crucial solutions for the specific problem of private inference, the contribution of ASTER is not to propose a new inference algorithm. Instead, our novelty lies in the holistic integration of a secure inference mechanism within a complete, end-to-end lifecycle framework.

A complementary approach is Differential Privacy (DP), which adds calibrated noise to model updates to provide formal privacy guarantees [9]. However, DP introduces a fundamental trade-off: stronger privacy requires more noise, which can significantly reduce model accuracy [10].

Finally, while general-purpose FL platforms like FATE, OpenFL, and PySyft provide toolkits for implementing these PETs, they are not prescriptive architectural solutions. In contrast, ASTER proposes a specific, integrated architecture that combines multiple cryptographic protocols by default. Furthermore, a key contribution of ASTER, rarely addressed in other frameworks, is the democratic governance of the model lifecycle. By incorporating SSS for controlling retraining, ASTER introduces a mechanism for distributed, threshold-based consensus, protecting the model from unauthorized unilateral modifications. This holistic approach, combining anonymity, confidential inference, and democratic control, addresses the need for a more robust and comprehensive solution for the entire FL lifecycle.

3. Multi-stage Privacy Framework

The proposed framework comprises three interconnected stages, each designed to protect a specific phase of the FL lifecycle against privacy breaches. The architecture ensures data protection during model training, inference, and the final aggregation process.

3.1. Anonymized Training Phase

The framework incorporates a mix-net architecture to address the risk of linking a user's data profile to their initial model contribution. This approach, inspired by anonymous electronic voting systems [11], decouples the user's identity from their trained model before it reaches the central server.

The process begins after a client trains a model on its local dataset and employs a layered encryption scheme. First, the client creates a secure package intended only for the final server: the model parameters are encrypted with a symmetric key (K_{sym}) via a cipher like AES, and this key is then separately encrypted with the central server's public key (PK_{Server}) via a scheme like ElGamal. Second, this package is encrypted again, this time using the public key of the intermediate server (PK_{Mixer}), creating an outer cryptographic layer (see Fig. 1).

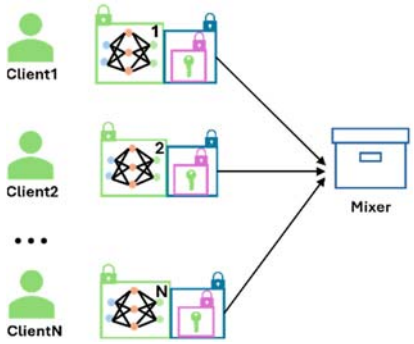


Fig. 1. Clients send layered encrypted packages.

This final double-encrypted package is submitted to an intermediate server, the Mixer. Upon receiving a batch of packages, the Mixer uses its private key (SK_{Mixer}) to decrypt only the outer layer, revealing the still-encrypted packages. The Mixer, therefore, has no access to the actual model parameters. It then shuffles these packages to break the relation between the incoming and outgoing order (illustrated in Fig. 2) and forwards the reordered batch to the central server.

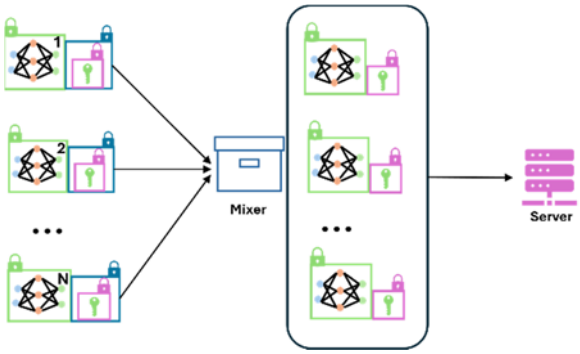


Fig. 2. Mixer decrypts and shuffles packages.

Consequently, the central server receives and decrypts a set of fully trained local models but is computationally prevented from determining the source of any individual model. The server then aggregates these anonymized models to produce the federated model, having preserved user anonymity throughout the process.

3.2. Secure Inference Phase

Once the federated model is established, clients require a secure method to use it for predictions in the central server without revealing their sensitive input data. The framework leverages HE to facilitate this private inference phase, with support for linear autoregressive models and tree-based models.

A client encrypts its input vector, x_i, with its HE public key, and sends the resulting ciphertext c_i to the

server. The server then processes this ciphertext using a homomorphic adapted global model. The server executes the inference pass directly on the encrypted data, producing an encrypted prediction, y_{enc}. This encrypted result is returned to the client. Finally, only the client can decrypt y_{enc} using their corresponding private key to obtain the final prediction in plaintext. This process guarantees that the server never accesses the client's input data or the final prediction, ensuring end-to-end confidentiality.

3.3. Secure Retraining Phase

To prevent unauthorized modifications and ensure that federated model updates result from a broad consensus, the framework implements a conditional aggregation mechanism based on a SSS. This adds a layer of democratic and robust control over the model's lifecycle [12].

A *(t,n)*-threshold Shamir's Secret Sharing scheme is employed, where n is the total number of clients in the federation and t is the minimum participation threshold required to authorize an update. The protected secret is a master key, K_M, which enables the next aggregation round. This key is split into n unique shares, $\{s_1,...,s_n\}$, and each client is securely issued one share. By design, any subset of fewer than t shares reveals no information about the master key.

When a new model aggregation round is proposed, clients who wish to participate must submit their shares to the server. If the number of submitted shares is greater than or equal to the threshold t, the master key K_M can be successfully reconstructed. The reconstructed key is then used to authorize and execute the aggregation process. This prevents unauthorized model modifications by small coalitions and enforces distributed, fault-tolerant control over the model's evolution.

4. Conclusions and Future Work

This paper introduced ASTER, a holistic framework providing end-to-end FL privacy by combining an e-voting inspired mix-net for anonymity, HE for private inference, and a SSS for conditional aggregation. The result is a system that protects user data while ensuring the integrity and democratic control of the global model's evolution.

Future work will focus on two key challenges. The first is extending HE-based secure inference to Deep Learning models. This is non-trivial due to the complex substitution required for non-linear activation functions and the potential impact on model accuracy [13]. The second challenge is the significant computational latency introduced by HE operations. To address this, an alternative architecture will be explored where clients securely download the global model to perform fast, local predictions on their own devices. This hybrid approach could offer the ideal balance between the privacy of server-side computation and the low-latency performance required for real-time applications.

References

[1]. N. Bouacida, P. Mohapatra, Vulnerabilities in federated learning, *IEEE Access*, Vol. 9, 2021, pp. 63229-63249.

[2]. K. Wei, J. Li, M. Ding, C. Ma, et al., Federated learning with differential privacy: algorithms and performance analysis, *IEEE Transactions on Information Forensics and Security*, Vol. 15, 2020, pp. 3454-3469.

[3]. H. Fang, Q. Qian, Privacy preserving machine learning with homomorphic encryption and federated learning, *Future Internet*, Vol. 13, Issue 4, 2021, 94.

[4]. R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*, 2017, pp. 3-18.

[5]. Y. Li, Y. Zhou, A. Jolfaei, D. Yu, et al., Privacy-preserving federated learning framework based on chained secure multiparty computing, *IEEE Internet of Things Journal*, Vol. 8, Issue 8, 2021, pp. 6178-6186.

[6]. H. Zhu, R. S. M. Goh, W. K. Ng, Privacy-preserving weighted federated learning within the secret sharing framework, *IEEE Access*, Vol. 8, 2020, pp. 198275-198284.

[7]. J. Ma, S. A. Naas, S. Sigg, X. Lyu, Privacy-preserving federated learning based on multi-key homomorphic encryption, *International Journal of Intelligent Systems*, Vol. 37, Issue 9, 2022, pp. 5880-5901.

[8]. C. Zhang, S. Li, J. Xia, W. Wang, et al., {BatchCrypt}: efficient homomorphic encryption for {Cross-Silo} federated learning, in *Proceedings of the USENIX Annual Technical Conference (USENIX ATC'20)*, 2020, pp. 493-506.

[9]. A. El Ouadrhiri, A. Abdelhadi, Differential privacy for deep and federated learning: a survey, *IEEE Access*, Vol. 10, 2022, pp. 22359-22380.

[10]. S. Truex, L. Liu, K. H. Chow, M. E. Gursoy, et al., LDP-Fed: federated learning with local differential privacy, in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking (EdgeSys'20)*, 2020, pp. 61-66.

[11]. Y. -X. Kho, S. -H. Heng, J. -J. Chin, A review of cryptographic electronic voting, *Symmetry*, Vol. 14, Issue 5, 2022, 858.

[12]. K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, et al., Practical secure aggregation for privacy-preserving machine learning, in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175-1191.

[13]. S. Obla, X. Gong, A. Aloufi, P. Hu, et al., Effective activation functions for homomorphic evaluation of deep neural networks, *IEEE Access*, Vol. 8, 2020, pp. 153098-153112.

(011)

A Guide to Feature-preserving Pseudonymization of Profile Pictures

Y. Lee, H. Bothe and M. Geierhos
University of the Bundeswehr Munich, RI CODE, Werner-Heisenberg-Weg 39, 85579 Munich, Germany
Tel.: +498960047343
E-mail: {yeongsu.lee, hendrik.bothe, michaela.geierhos}@unibw.de

Summary: To address the privacy risks associated with biometric data, we propose a method of pseudonymizing social media profile pictures. Unlike existing approaches that focus on text and structured data, profile images can directly identify users. Our pipeline uses FaceNet and DeepFace to extract facial attributes, such as age, gender, and expression. It then formats these attributes as JSON and converts them to descriptive text using Mistral (7B). A multimodal model called Janus then uses this text to generate synthetic, identity-free profile images that retain facial features. All processing is done locally to ensure compliance with the GDPR and avoid data exposure. These pseudonymized images support research and machine learning tasks while protecting privacy. To evaluate image quality and privacy, we cluster the original and pseudonymized images and compare them using the Adjusted Rand Index and Normalized Mutual Information metrics. These metrics assess semantic consistency and identity separation. Our method securely and ethically analyzes social media data by combining facial de-identification, large language models, and generative image synthesis in a unified workflow.

Keywords: Privacy of profile images, Facial feature extraction, Synthetic image generation.

1. Introduction

Profile pictures on social media platforms contain biometric information that can be used to directly identify individuals. Therefore, these images must be fully protected. Textual pseudonymization alone is insufficient. Since facial images can be cross-referenced across platforms or used for facial recognition, retaining real profile pictures in research datasets poses significant re-identification risks.

To address this issue, we are expanding the scope of conventional pseudonymization to include visual elements. Our proposed method replaces original face images with synthetic, anonymized alternatives that are semantically consistent. These synthetic images will retain attributes such as age, gender, ethnicity, and other facial features, including emotional expression. This allows for meaningful analysis while ensuring that individuals cannot be re-identified. Our approach improves upon earlier text- and attribute-based pseudonymization techniques by integrating local image synthesis in a privacy-safe manner. Our implementation operates entirely on local hardware, eliminating reliance on external APIs or cloud-based models. This reduces exposure to external threats and strengthens overall data security for sensitive research.

2. Related Work

The issue of face pseudonymization has been addressed in several domains. One notable approach is DeepPrivacy, which uses conditional generative adversarial networks (GANs) to synthesize faces while preserving the scene's context. This method allows for anonymization without cropping or blurring [1]. Building on this foundation, DeepPrivacy2 incorporates facial attributes such as age and gender while maintaining realism and enhancing control [2]. Another approach is IdentityDP [3], which introduces privacy-preserving transformations using differential privacy. These methods inject calibrated noise into face representations, which limits the risk of re-identification, even when adversaries have more data.

Recent studies have proposed anonymization frameworks that optimize the generative model alongside identity supervision components. These frameworks use adversarial identity constraints and attribute similarity losses to balance privacy and utility [4, 5]. However, few approaches integrate large language models (LLMs) into the pipeline. Building on this research, our method uses attribute extraction and LLM-based description generation to synthesize pseudonymous images with the multimodal Janus generative model. This allows for fine-grained control over the appearance of the synthesized face and enables flexible adaptation to different demographic needs. Unlike black-box GANs, our method is interpretable, extendable, and capable of local execution. It provides privacy without compromising analytical usefulness.

3. Our Approach

Our profile picture pseudonymization pipeline consists of five main stages from face detection to pseudonymized profile image generation.

(1) The first steps in the process are face detection and cropping. We use a FaceNet-based facial detection algorithm to extract faces from profile pictures, which

are marked by a red box that includes the detection score[1]. Then, we extend the detected area to include more information, such as hairstyle and accessories, as marked by the orange box in Fig. 1.

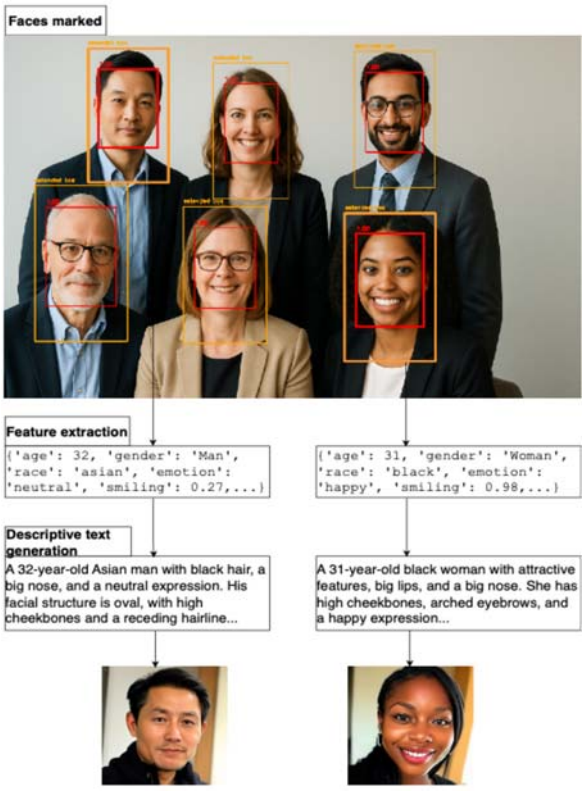


Fig. 1. An example of the pseudonymization process.

(2) Next, the extended images are passed to the facial attribute extraction tools. Using FaceNet [6] and DeepFace [7, 8], we identify features such as age, gender, emotion, ethnicity, skin tone, and head pose. These attributes are then structured into a standardized JSON format. There are 52 extracted attributes, categorized into 12 feature classes. These classes range from general features, such as gender and age, to emotions, including happy, sad, angry, and neutral.

(3) Then, the JSON-encoded attributes are sent as instructions to Mistral (7B), which is running locally through the Ollama interface[2]. This step is required because Janus does not seem to rightly interpreted the JSON-encoded data. In our experiment, it produced an abstract image that was irrelevant to the human profile image. The model produces coherent, natural-language descriptions of the face, such as "a smiling woman in her mid-30s with East Asian features and medium-length black hair". To coherently convert the JSON-encoded attributes, they are classified according to their inherent characteristics. Thus, while some attribute classes such as accessories including "eyeglasses", "wearing hat", "wearing necktie", "wearing necklace", or "wearing earrings" are encoded as Boolean values, other attributes classes such as "oval face" or "big nose" are encoded as gradient values. Other classes, such as "wearing lipstick", are encoded as scalable values. For example, we divide the attribute class "wearing lipstick" into five descriptive expressions according to the extracted values: no lipstick, subtle lipstick, moderate lipstick, noticeable lipstick, and bold lipstick. Both these expressions and the JSON-encoded data are fed to Mistral for the generation of an accurate text description of the original profile picture. For all attribute classes, the threshold is set according to the statistical calculation from the analyzed data.

(4) The descriptive text is input into Janus-Pro (7B)[3], a generative model that can produce realistic images of faces based on textual prompts. Janus synthesizes faces that preserve the extracted attributes while completely breaking the visual link to the original image. This output image replaces the original profile picture in the pseudonymized dataset.

(5) The new synthetic image is stored with the rest of the pseudonymized profile data. Fig. 1 shows an example of the main steps.

Therefore, this method replaces biometric identifiers with synthetic alternatives, protecting against facial recognition attacks and compliance violations. It allows for the use of downstream applications such as demographic studies, personality inference, and expression analysis without revealing real identities. Furthermore, the pipeline can be expanded to include other image generators, multilingual descriptions, and LLM upgrades (e.g., Llama 4) to enhance accuracy. Additionally, the proposed pipeline can be applied repeatedly to generate further profile images as shown in Fig. 2.

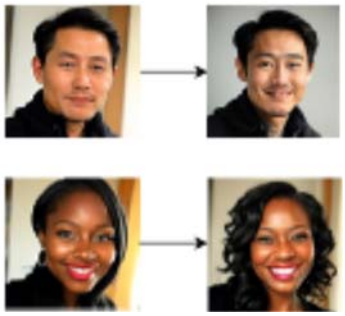


Fig. 2. An example of the recursive generation of pseudo-profile images.

4. Dataset and Ethical Consideration

The dataset used in this study consists of profile images obtained from publicly accessible XING profiles. Data of low quality (face detection confidence score < 0.95) was excluded from both statistics and experiments. Facial representations were extracted

[1] https://github.com/faustomorales/keras-facenet
[2] https://ollama.com/library/mistral
[3] https://github.com/deepseek-ai/Janus

using two state-of-the-art frameworks: **FaceNet**, yielding 40 numerical features[1], and **DeepFace**, providing 4 high-level categories comprising of 15 detailed features in total[2]. After consolidating the features from both frameworks, 52 features were categorized and used for analysis.

To understand the distributional characteristics of the dataset, we conducted an analysis of **demographic attributes** inferred from the facial representations, including **estimated age, gender, and ethnicity**. These estimates were obtained via DeepFace's pre-trained demographic classifiers. While such estimations are inherently approximate and potentially biased, they were used solely to assess diversity within the dataset, not to draw conclusions about individuals. Our demographic analysis revealed a heterogeneous distribution of age groups and gender identities, although the dataset may exhibit platform-specific biases. No manual annotation of demographic traits was performed, and no sensitive or protected categories were directly labeled or inferred beyond algorithmic estimation.

We acknowledge the ethical complexity of working with facial and demographic data, particularly regarding the risks of **profiling, bias, and unintended inferences**. To mitigate these risks: (1) only publicly available data was used, (2) all features were processed locally, and (3) no models were trained or evaluated for personal identity or real-world demographic classification tasks. Table 1 shows the distribution of gender and ethnicity of the used dataset based on the classification results of DeepFace. Please note that the number of genders and ethnicities is limited to the capabilities of DeepFace.

Table 1. Demographic statistics of the used dataset.

	Men	**Women**
White	14,524	4,426
Latino-Hispanic	1,002	206
Middle Eastern	760	24
Asian	503	145
Indian	134	6
Black	84	10
Total	**17,007**	**4,817**

5. Discussion

To evaluate the effectiveness of profile picture pseudonymization, we must analyze both privacy protection and data utility. We propose using clustering-based evaluation methods. First, a consistent face embedding model is used to generate facial embeddings for the original and pseudonymized images. Then, clustering is applied to both sets. We assess the similarity of the clustering structures using the adjusted rand index (ARI) and the normalized mutual information (NMI). High NMI and ARI values suggest that key semantic attributes, such as gender or age, are preserved. Conversely, a low identity match across clusters confirms successful anonymization. Additional validation can be conducted using face-matching algorithms to confirm that no identity overlap remains. Human evaluation can also be used to assess realism and anonymity. Other metrics, such as Fréchet inception distance (FID), can be used to measure perceptual realism. However, limitations include biases in attribute extraction, generation artifacts, and ambiguity in prompt interpretation. Future work should include automated bias detection and adversarial robustness testing.

6. Conclusion

This work expands the scope of pseudonymization to include visual content, specifically profile pictures, in addition to structured and textual data. Our method achieves privacy-preserving replacement of facial identity by extracting facial attributes, transforming them via LLM-generated text, and synthesizing new images with Janus. Evaluation via clustering and identity metrics ensures that the resulting images are both anonymized and analytically useful. This pipeline adheres to privacy-by-design principles and can be fully executed locally to minimize data leakage risks.

Acknowledgements

We thank dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

References

[1]. H. Hukkelås, R. Mester, F. Lindseth, DeepPrivacy: a generative adversarial network for face anonymization, in Advances in Visual Computing (G. Bebis, et al., Eds.), *Springer*, 2019, pp. 565-578.

[2]. H. Hukkelås, F. Lindseth, DeepPrivacy2: towards realistic full body anonymization, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'23)*, 2023, pp. 1329-1338.

[3]. Y. Wen, B. Liu, M. Ding, R. Xie, et al., IdentityDP: differential private identification protection for face images, *Neurocomputing*, Vol. 501, 2022, pp. 433-447.

[4]. M. Maximov, I. Elezi, L. Leal-Taixé, CIA-GAN: conditional identity anonymization generative adversarial networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*, 2020, pp. 5446-5455.

[5]. S. Barattin, C. Tzelepis, I. Patras, N. Sebe, Attribute preserving face dataset anonymization via latent code optimization, in *Proceedings of the IEEE/CVF*

[1] https://tinyurl.com/26qp7jem

[2] https://github.com/serengil/deepface

Conference on Computer Vision and Pattern Recognition (CVPR'23), 2023, pp. 8316-8326.

[6]. F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015, pp. 815-823.

[7]. Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: closing the gap to human-level performance in face verification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*, 2014, pp. 1701-1708.

[8]. S. Serengil, A. Ozpinar, A benchmark of facial recognition pipelines and co-usability performances of modules, *Journal of Information Technologies*, Vol. 17, Issue 2, 2024, pp. 95-107.

(012)

A Microservice Based Authentication and Authorization Framework for Column Oriented Databases

Rafael Sebastian Castro Paredes, Andres Andrade-Cabrera and Diana Martinez-Mosquera
[1] Escuela Politécnica Nacional, Department of Informatics and Computer Science, Quito, Ecuador
E-mail: {rafael.castro, luis.andrade03, diana.martinez}@epn.edu.ec

Summary: Modern systems often combine transactional and analytical databases, forming polyglot persistence architectures that pose serious challenges for authentication and access control. NoSQL databases, like Apache Cassandra, dominate in analytical use cases but lack native security features. A common workaround is to rely on separate relational databases for authentication, which creates bottlenecks and complexity. This paper presents a microservice-based security framework designed specifically for column-oriented NoSQL environments. It includes dedicated microservices for user registration, secure credential validation using Bcrypt, and issuing JSON Web Tokens (JWT) for authentication. Authorization is enforced via a separate service implementing fine-grained, role-based access control (RBAC). The system is implemented using NestJS (backend), React (frontend), and Cassandra as the unified datastore for users and permissions. All components are containerized with Docker for scalability and deployment ease. Experimental results show that the framework improves security separation, enhances flexibility, and facilitates secure adoption of NoSQL in distributed systems.

Keywords: Access control, Authentication, Cassandra, Microservices, NoSQL, Security framework.

1. Introduction

In the era of rapid data growth, organizations are moving away from relying on a single database solution, embracing polyglot persistence environments instead [1]. This approach leverages different types of databases such as SQL, NoSQL, and NewSQL [10] to meet the specific requirements of various components within a company. Among NoSQL options, column-oriented databases are widely adopted due to their high availability and excellent scalability, making them ideal for processing large volumes of data [1].

However, column-oriented databases were not designed with built-in authentication and authorization mechanisms [11, 12], which poses challenges for developers and database administrators. Centralized security models are poorly suited to the distributed nature of modern applications. This research addresses the design of a flexible and secure authentication and authorization system tailored for column-oriented databases. The proposed system must reliably verify user identities and enforce fine-grained access control, ensuring that users can only perform actions they are explicitly authorized to execute.

This paper contributes a few key things. First, we propose a decoupled security architecture made of specialized microservices for authentication and permissions. This proposal improves how modular and scalable the system is. Second, this study details a secure authentication process using JSON Web Tokens (JWT) [13] for session management and Bcrypt [14] for password hashing. Third, we present a dynamic model for access control based on roles. This approach lets administrators manage user permissions for database operations in real time.

The rest of this paper is structured to explain our work. Section 2 looks at the related research. Section 3 describes the system's architecture. Section 4 explains the implementation details of the proposal. Section 5 shows the evaluation results. Finally, Section 6 concludes the paper with thoughts on future work.

2. Related Work

Security is very important in distributed systems. Actually, the microservices architectural pattern [2, 3] is a standard way to build scalable applications. The main idea of this pattern is breaking up responsibilities, which includes security. Building authentication and authorization as a dedicated microservice is a best-established practice. This approach prevents concentrating all security responsibilities within a single Application Programming Interface (API) gateway, while enabling security logic to be developed, maintained, and scaled independently.

Securing column-oriented databases requires a combination of traditional access control, privilege management, and innovative techniques like shuffling and encryption. As data models become more complex, authorization frameworks must evolve to ensure both security and usability without compromising performance [4].

Unlike other works [5-9], which introduces a theoretical authorization model for object-oriented and semantic databases with a focus on implicit authorization, our proposal addresses the practical challenges of securing polyglot persistence environment, specifically column-oriented NoSQL databases. While their approach extends traditional models conceptually within single-system architectures [5-9], our framework provides a modular, microservice-based solution for both authentication and fine-grained authorization, using modern tools such as JWT, Bcrypt, and Docker [15]. This makes our system suitable for real-world, distributed applications

that require scalable and flexible access control mechanisms.

3. Methodology

The proposed architecture is based on a microservice-based security layer and introduces a modular framework designed specifically for a well-known column-oriented NoSQL database, such as Apache Cassandra [16]. This solution aimed at managing authentication, authorization, and query mediation.

Fig. 1 depicts the overall structure of the system, including the main components and how they are connected. The architecture includes a frontend client built with React, three backend microservices built with NestJS: (1) Translation Microservice, (2) Authentication Microservice, and (3) Permissions Microservice; and a Cassandra database cluster running in Docker.

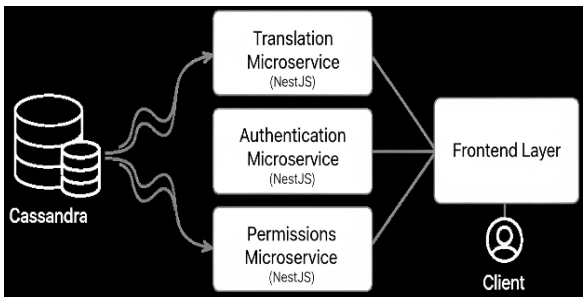


Fig. 1. Microservice Security Architecture for Cassandra.

The Translation Microservice is the main engine that translates SQL queries to Cassandra Query Language (CQL) and runs them after checking permissions. Authentication Microservice is the main gatekeeper that handles user registration, login, and password management. The Permissions Microservice manages all authorization rules and defines what actions a user can take.

The frontend is a single page application, which provides the user interface for login, registration, queries, and admin tasks. Cassandra fulfills two roles: (1) it serves as the target database for user queries, and (2) it stores the application's internal data, such as user credentials, roles, and permission within a special auth keyspace.

3.1. Backend Microservices

3.1.1. Translation Microservice

We have incorporated a module of Translation that enables seamless interaction between applications using SQL and the Cassandra database. Its primary role is to translate incoming SQL queries into their equivalent Cassandra Query Language (CQL) statements. This approach provides two major benefits:

1. Compatibility: Applications or users accustomed to SQL can interact with Cassandra without needing to learn CQL syntax directly;
2. Abstraction: Developers can focus on business logic rather than query translation details, improving productivity and reducing errors.

3.1.2. Authentication Microservice

This service confirms a user's identity. Its main functions are user registration, login, password recovery, and other security mechanisms. User registration creates new accounts. It hashes passwords with Bcrypt to avoid storing them in plain text. When a user registers, the system generates a unique PIN for password recovery.

Login validates user credentials, if they are correct, it generates a signed JWT with the user's ID and role, like USER or ADMIN. Password recovery gives users a secure way to reset their password. Users can use their original PIN or a temporary one from an administrator. Security measures include rate limiting, which mitigates brute-force attacks by temporarily blocking login attempts after a predefined number of failures.

3.1.3. Permissions Microservice and Role-based Access Control Model

The Permissions Microservice handles authorization and uses a model for access control named Role-Based Access Control (RBAC). This model defines two main roles:

1. User: A regular account with permission to run defined CQL queries on authorized Cassandra keyspaces;
2. Admin: A privileged account with full access to all translation functions and can manage the system.

Admin tasks include managing user roles, assigning access to keyspaces, granting permissions for specific operations, generating temporary PINs, and deleting users or keyspaces. Permission data is stored directly in Cassandra, enabling dynamic and persistent control over user actions.

3.2. Authentication and Authorization Flow

A user first logs in through the frontend by sending their credentials to Authentication Microservice. The service checks the credentials and, if they are correct, sends back a JWT. The frontend saves this JWT and adds it to the header of all future requests. When the user tries to run a query, a JwtAuthGuard checks the JWT's signature and expiration date.

Before running the query, the Translation Microservice asks the Permissions Microservice if the

user is allowed to perform that action on that keyspace. The Permissions Microservice checks its database and responds with approval or denial. The Translation Microservice will only proceed if it gets approval.

Fig. 2 illustrates the step-by-step process of how a user is authenticated and authorized to perform actions within the system.

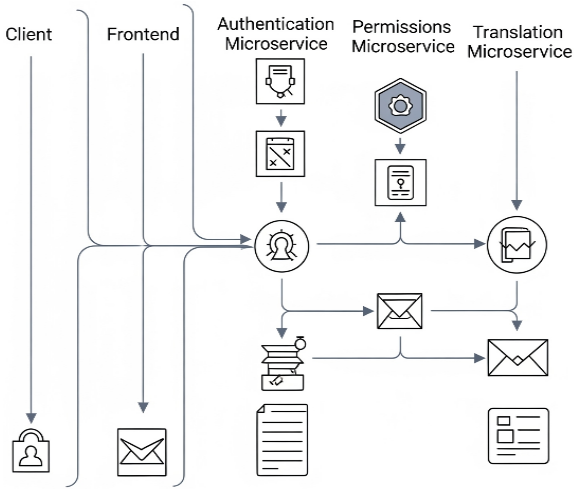


Fig. 2. Authentication and Authorization Flow.

3.3. Implementation

Fig. 3 depicts the implementation of the backend and frontend. The backend was built with NestJS. This is a Node.js framework for building scalable server-side applications. Its modular design was a good fit for our microservice approach. It is containerized with Docker for simple deployment and consistency. The complete source code and deployment instructions are publicly available on GitHub to facilitate reproducibility [17].

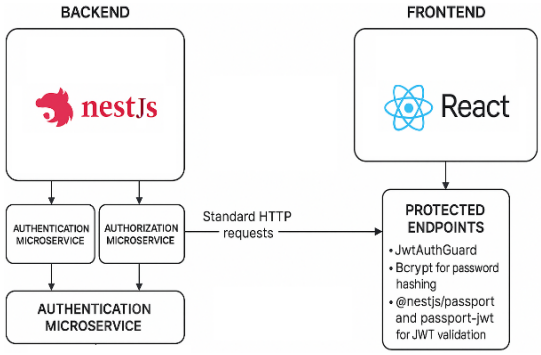


Fig. 3. Implementation of the Architecture for Authentication and Authorization.

The frontend was built with React and Vite, this one created a fast and responsive user interface. For security, we used the jsonwebtoken library to create and check JWTs. We used Bcrypt for one way hashing of passwords. We also used @nestjs/passport and passportjwt to integrate JWT validation into the application's request process using custom Guards.

The Authentication Microservice provides REST endpoints for login, registration, and password changes. The Permissions Microservice provides admin endpoints for updating user permissions and keyspaces. All sensitive endpoints are protected by a JwtAuthGuard that checks the token and a RolesGuard that checks if the user has the required role to access the resource. The services communicate with each other using standard HTTP requests.

3.4. Performance Evaluation

To validate the scalability and robustness of the proposed framework, a series of performance tests were conducted. These tests move beyond the initial functional validation to provide empirical benchmarks for latency, throughput, and system resilience under significant load, addressing the future work outlined in our initial draft.

The evaluation was performed across two distinct environments to isolate and accurately measure the performance of the microservice architecture:

1. Environment A (Local Development): Initial baseline tests were conducted on a developer laptop equipped with a multi-core processor and 16GB of RAM. This environment was used for iterative functional testing and to establish a preliminary performance baseline under light load (1-10 concurrent users);
2. Environment B (Dedicated Server): The formal scalability and stress tests were executed on a dedicated Linux server featuring a 64-core processor and 32GB of RAM. This powerful environment ensured that test results reflect the true performance of the application architecture, free from potential bottlenecks of a local machine.

Furthermore, we utilized the open-source tool k6 to simulate concurrent user traffic and measure key performance indicators. Two primary scenarios were designed:

1. Scalability Test: This test measured the system's ability to handle a linearly increasing load. The number of virtual users (VUs) was ramped up in stages from 1 to 200 over a 16-minute period. The primary metrics were requesting throughput (requests/second), P95 latency, and the error rate;
2. Stress Test: This test was designed to determine the system's stability and breaking point under extreme load. The number of VUs was aggressively ramped up to 500 over an 11-minute period to saturate the services and observe their behavior under maximum stress.

4. Results and Discussion

The system was tested with a series of functional tests. These tests covered all important security and

administrative workflows. The results confirmed that the framework was implemented correctly.

The authentication tests showed that the login interface successfully logged in users with valid credentials and blocked invalid ones. The user registration process correctly created new users and gave them a unique PIN for recovery.

The authorization tests showed that the admin panel for permissions allowed an admin to enable or disable specific SQL operations for a user. The backend logs confirmed these changes were saved to the database and sent to the translation service to update its cache.

The role management tests showed the system correctly handled promoting a user to an admin. The backend log showed the successful role update. After the promotion, the user could access the administrative dashboard.

The password recovery tests were also successful. Both the initial PIN and a temporary admin generated PIN worked to reset passwords through the interface. These results show that architecture based on microservices provides a solid and functional security layer for the Cassandra database.

Furthermore, the system's performance remained stable, and all queries were executed successfully without errors or noticeable delays.

4.1. Graphical User Interface

Fig. 4 depicts a login interface for Cassandra NoSQL database. On the left side of the image is the login form, which includes fields for Username, ID/Code, and Password, along with options to Register or Recover a forgotten password. CASSQL is a platform that allows users to authenticate and interact with Cassandra using SQL-like inputs, offering a user-friendly interface for data access and translation.

Fig. 4. Authentication GUI.

At the top, there are input fields to **search users by ID**, and assign or view their **Name**, **ID/Code**, and **Role**. Below, a grid of permissions is displayed, allowing fine-grained control over actions like:

- **Schema-level operations**: Create Table, Drop Table, Alter Table (Add, Drop, Rename), Create Keyspace, Drop Keyspace;
- **Data-level operations**: Select, Insert, Update, Delete, Truncate;
- **Metadata operations**: Describe Table(s), Create/Drop Index, Alter Keyspace.

Fig. 5 shows the "Permissions Configuration" interface of the CASSQL platform. This module allows an administrator to manage user access rights for operations in a Cassandra-based environment. At the top, there are input fields to search users by ID, and assign or view their Name, ID/Code, and Role. Below, a grid of permissions is displayed, allowing fine-grained control over actions like:

1. Schema-level operations: Create Table, Drop Table, Alter Table (Add, Drop, Rename), Create Keyspace, Drop Keyspace;
2. Data-level operations: Select, Insert, Update, Delete, Truncate;
3. Metadata operations: Describe Table(s), Create/Drop Index, Alter Keyspace;
4. Access control: Use (general access).

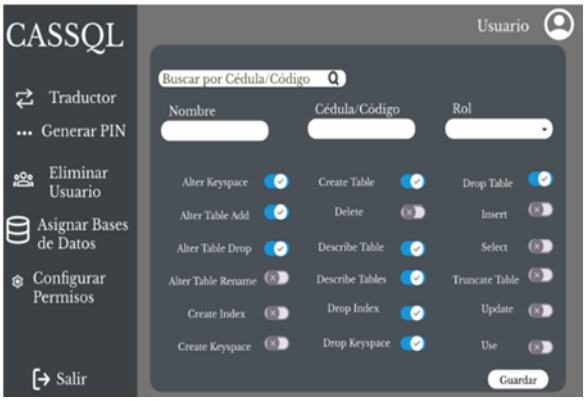


Fig. 5. Authorization GUI.

Each permission can be toggled on or off using a switch. Once the desired permissions are configured, the “Save” button at the bottom right applies the changes.

The left sidebar includes navigation options such as Translator, Generate PIN, Delete User, Assign Databases, and Configure Permissions, making it a full-featured role and access management interface for Cassandra.

4.2. Performance

The results from the performance evaluation were exceptionally positive, demonstrating that the microservice architecture is both highly scalable and resilient.

4.2.1. Scalability

Under the scaling load of 200 concurrent users, the system stabilized at a high throughput of 141.6 requests per second. The 95th percentile (p95) latency for database operations remained low, with

INSERT operations at 335 ms and SELECT operations at 369 ms. Critically, the system maintained a 0.00 % error rate throughout the entire test.

As shown in Fig. 6, the P95 latency for both INSERT and SELECT operations increases linearly and predictably as the number of concurrent users scales to 200. This controlled rise in latency, coupled with the 0 % error rate, is indicative of a healthy and scalable system that gracefully handles increasing demand without becoming unstable.

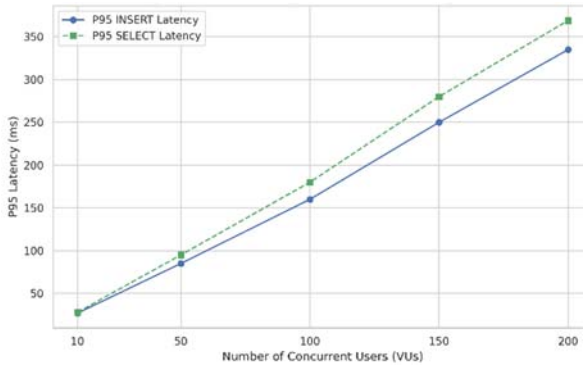


Fig. 6. P95 Latency during Scalability Test.

4.2.2. Stress Resilience

During the stress test with 500 concurrent users, the system demonstrated remarkable stability. It processed a peak throughput of 338.2 requests per second. While p95 latency increased to approximately 778 ms under this extreme load, the system never failed. The error rate remained at 0.00 %, proving that the architecture is resilient and does not buckle under pressure that far exceeds typical operational loads.

Fig. 7 provides a clear comparison of the average and 95th percentile latency during the peak of the stress test. While the P95 latency is notably higher than the average, as expected under heavy load, both metrics remain within reasonable bounds for a system under such extreme conditions. This reinforces the finding that the system is resilient and does not exhibit uncontrolled latency spikes, even when pushed far beyond its expected operational capacity.

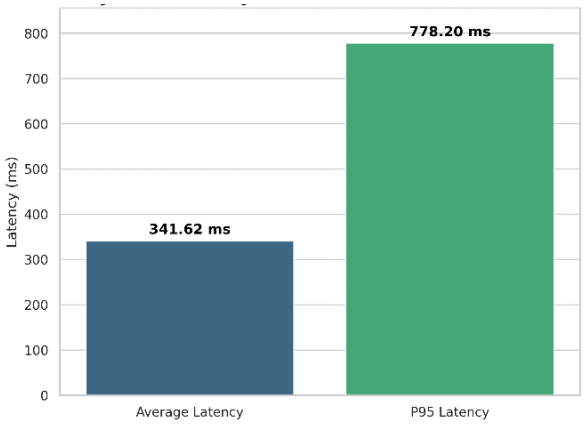


Fig. 7. Latency Comparison under Peak Stress Test.

These empirical results validate that the chosen microservice-based design, which decouples authentication, permissions, and translation, does not introduce performance bottlenecks. Instead, it provides a robust foundation capable of serving high-traffic, production-level environments.

5. Conclusions

This paper presented a comprehensive authentication and authorization framework tailored for column-oriented databases such as Cassandra. By leveraging microservices architecture, security concerns are clearly separated from core database operations. This separation enables the system to be highly scalable, maintainable, and resilient, while reducing complexity in both development and deployment.

The adoption of modern security standards, including JWT for secure session management and Bcrypt for one-way password hashing, ensures robust protection against common attack vectors. Furthermore, the integration of RBAC model introduces flexible and fine-grained control over data access, strengthening both user-level and resource-level security.

The successful design, implementation, and testing of this framework demonstrate its viability as a reliable solution for securing complex NoSQL database systems. By lowering the security barrier, it also makes powerful distributed technologies such as Cassandra more accessible to developers, encouraging safer adoption in enterprise and research environments.

Performance evaluation confirms the robustness of the proposed system. It maintained high throughput, low latency, and a 0 % error rate under normal and stress conditions. Even with 500 concurrent users, the system remained stable and resilient, showing predictable performance without failures or uncontrolled spikes.

Looking forward, several future enhancements could strengthen and extend this framework. First, incorporating multi-factor authentication (MFA) would significantly enhance login security by adding an additional verification layer. Second, integrating support for widely used identity and federation protocols, such as OAuth 2.0 and OpenID Connect, would enable seamless authentication with trusted third-party providers. Finally, conducting a more in-depth security threat analysis, covering scenarios such as token hijacking, replay attacks, or privilege escalation, would allow further refinement of the framework and proactive mitigation of emerging threats.

References

[1]. J. Carpenter, E. Hewitt, Cassandra: The Definitive Guide, 2nd Ed., *O'Reilly Media*, 2016.

[2]. IBM, Microservices, https://www.ibm.com/es-es/topics/microservices

[3]. E. Wolff, Microservices: Architecture and Design, *Addison-Wesley*, 2016.

[4]. T. Geng, C. -T. Huang, C. Farkas, SCORD: shuffling column-oriented relational database to enhance security, in *Proceedings of Ubiquitous Security (UbiSec'23)*, 2023, pp. 163-176.

[5]. F. Rabitti, E. Bertino, W. Kim, D. Woelk, A model of authorization for next-generation database systems, *ACM Transactions on Database Systems*, Vol. 16, Issue 1, 1991, pp. 88-131.

[6]. A. Mohamed, D. Auer, D. Hofer, J. Küng, A systematic literature review of authorization and access control requirements and current state of the art for different database models, *International Journal of Web Information Systems*, Vol. 20, 2024, pp. 1-23.

[7]. S. Chaudhuri, T. Dutta, S. Sudarshan, Fine grained authorization through predicated grants, in *Proceedings of the IEEE 23rd International Conference on Data Engineering*, 2007, pp. 1174-1183.

[8]. G. Deep, R. Mohana, A. Nayyar, S. Padmanaban, et al., Authentication protocol for cloud databases using blockchain mechanism, *Sensors*, Vol. 19, 2019, 4444.

[9]. M. Jiang, S. Liu, S. Han, D. Gu, Biometric-based two-factor authentication scheme under database leakage, *Theoretical Computer Science*, Vol. 1000, 2024, 114552.

[10]. T. Khasawneh, M. Al-Sahlee, A. Safia, SQL, NewSQL, and NoSQL databases: a comparative survey, in *Proceedings of the 11th International Conference on Information and Communication Systems (ICICS'20)*, 2020, pp. 13-21.

[11]. D. Abadi, P. Boncz, S. Harizopoulos, Column oriented database systems, *Proceedings of the VLDB Endowment*, Vol. 2, Issue 2, 2009, pp. 1664-1665.

[12]. I. Solsol, H. Vargas, G. Díaz, Security mechanisms in NoSQL DBMS's: a technical review, in Advances in Intelligent Systems and Computing, Vol. 1362, *Springer*, 2021, pp. 215-228.

[13]. M. Jones, J. Bradley, N. Sakimura, JSON web token (JWT), RFC 7519, *IETF*, 2015, pp. 1-30.

[14]. C. Skanda, B. Srivatsa, B. Premananda, Secure hashing using BCrypt for cryptographic applications, in *Proceedings of the IEEE North Karnataka Subsection Flagship International Conference (NKCon'22)*, 2022, pp. 1-5.

[15]. M. Yasir, A review on introduction to Docker and its features, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 8, Issue 6, 2018, pp. 104-108.

[16]. H. Tu, Cassandra vs. MongoDB: a systematic review of two NoSQL data stores in their industry uses, in *Proceedings of the IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI'24)*, 2024, pp. 81-86.

[17]. R. Castro, Middleware-NoSQL/SQL-to-CQL, GitHub Repository, https://github.com/Middleware-NoSQL/SQL-to-CQL

(013)

Developing Synthetic Data or Cybersecurity Policies

G. Giacomello and O. Preka
Center for Computational Social Science, DSPS, University of Bologna, Strada Maggiore 45, Bologna, Italy
Email: giampiero.giacomello@unibo.it

Summary: Data scarcity – driven by under-detection and under-reporting of incidents, legal and competitive disincentives to share, and vendors' reluctance to expose product weaknesses – impedes the development of data-driven cybersecurity policies. We investigate large language model (LLM)–based synthetic tabular data generation as a pragmatic remedy. Our approach follows a GReaT-style pipeline: (i) text-based serialization of heterogeneous tables to preserve schema semantics; and (ii) fine-tuning a pretrained decoder LLM (Unsloth/Llama-3.2-1B) on 3388 publicly reported cyber-attack records. Preliminary results that LLM-generated synthetic data can approximate the statistical and structural properties of scarce cybersecurity data without exposing sensitive information, thereby enabling data augmentation and supporting the design of data-driven cyber policies.

Keywords: Synthetic data, Cybersecurity, Tabular data generation.

1. Research Problem

The development of data-driven cybersecurity policies and defensive mechanisms is critically hampered by a persistent, multifaceted challenge: data scarcity. In an era where threats are increasingly sophisticated, the reliance on comprehensive datasets for training intrusion detection systems, analysing malware behaviour, and modelling threat landscapes has never been greater. However, data availability is extremely scarce for several reasons. Key factors include the small fraction of cyber-attacks being detected and even less reported. Existing datasets suffer from incompleteness and inconsistency, thus raising significant concerns regarding their reliability and validity.

This data scarcity is not accidental but systemic, stemming from several factors. First, the very nature of cyber-attacks involves malicious actors who actively conceal their identity and methods, making data collection inherently difficult. Second, organizations are reluctant to share incident data due to legitimate fears of exposing unpatched vulnerabilities, incurring substantial reputational damage, and creating competitive disadvantages. This hesitation is further compounded by a complex web of legal and regulatory implications surrounding data privacy and breach notifications. Finally, technology vendors and software companies are often constrained by market pressures, making them hesitant to reveal product weaknesses that could be inferred from detailed attack data. These constraints create a critical bottleneck that impedes the research and development of data-driven cybersecurity solutions.

2. Methodological Approach: Synthetic Data Generation with LLMs

To address these significant challenges, we explore the potential of synthetic data generation, specifically through the application of Large Language Models (LLMs) as a novel solution. Synthetic data, defined as algorithmically generated information that resembles the statistical properties and structural characteristics of a source dataset, offers a powerful paradigm for overcoming data access barriers. Its utility has been demonstrated in various domains for tasks such as data anonymization to protect privacy in sensitive sectors like healthcare, data augmentation to enrich sparse datasets, and bias mitigation in machine learning models.

LLMs, trained on vast quantities of text, code, and other data types, have shown advanced generative capabilities and versatile problem-solving skills that extend well beyond traditional Natural Language Processing [1]. Their abilities such as in-context learning and instruction following can be leveraged via sophisticated prompt engineering techniques [2]. This has opened up opportunities for new applications of LLMs, including the generation of tabular data. While the application of LLMs to generate structured tabular data is a nascent field, recent advancements have shown promising results in areas such as imputing missing values [3], augmenting sparse data [4], and rebalancing imbalanced classes [5], laying the foundational for our investigation into synthetic data generation [6].

3. Methodology and Experimental Design

Our methodology leverages state-of-the-art LLM-based frameworks for tabular data generation. Recognising the inherent challenges of tabular data – which includes heterogeneous feature types (e.g., categorical, numerical), class imbalance across categories [5], and complex context-based interconnections [7] – we begin with a key preprocessing step: text-based serialization. This process converts structured tabular data into a linear textual representation that can be processed by an

LLM, translating the table's schema and values into a coherent sequence while preserving semantic knowledge from column names [8].

We then use prompt engineering to steer the generative process. This includes clear task instructions, utilizing in-context learning with few-shot examples for format and schema adherence, and step-by-step reasoning with Chain-of-Thought (CoT) prompts [9] where appropriate.

Current state-of-the-art methods for tabular data generation with large language models generally follow the same approach, though with some variations. **GReaT** [10] fine-runes a pretrained GPT-2 model on the original data, and employs random permutations of feature order to reduce dependence on column sequences. For more complex scenarios, **ReaLTabFormer** [11] is noted for its enhanced performance on relational databases. Other prominent methods include **TAPTAP** [12], which pretrains GPT-2 on a vast collection of public tables to build foundational knowledge, and **Tabula** [13], which accelerates training by starting from a randomly initialized model.

Our experimental work focuses on the implementation of the **GReaT** approach, due to its advantages compared to other modes consisting mainly in a straightforward implementation without significant data processing. We apply the Great method on a collection of several publicly available reported cyber attacks, with the final dataset containing 3388 cases. We focus particularly on the technical characteristic (i.e., source name, target ID and target name), and free-text descriptions.

After an 80/20 train-test split, **Unloth/Llama-3.2-1B**, a pretrained transformers-decoder LLM [14] is fine-tuned on the train data set.

The evaluation of this method focuses both on fidelity and privacy of synthetic data: how close generated data is compared to unseen record from the original dataset, still without memorizing them as a result of overfitting. The most appropriate measure for this purpose is the Distance to Closest Record (DCR) measure [15]. For each data point, it calculates the distance to its nearest neighbor. Fig. 1 displays the distribution of minimal distances between data sets, showing that the generated data are close to the test data set, yet not exactly the same.

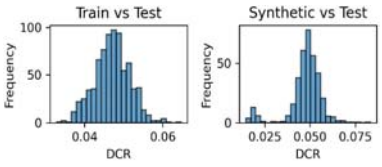


Fig. 1. DCR distributions computed with standardized numerical features and L2-normalized LM embeddings, using Euclidean (cosine-equivalent) distance. “Train vs Test” shows the DCR distance between training and test data set, while “Synthetic vs Test” shows the distance between synthetic data and test data. The synthetic distribution is slightly right-shifted (larger DCRs), indicating fewer near-duplicates of test data while maintaining comparable realism through overlapping ranges.

4. Conclusion

We demonstrate the feasibility of generating synthetic tabular data in the cybersecurity domain with an LLM-based approach adapted from GReaT and fine-tuned on a categorical-heavy corpus of nearly 3400 incidents.

Preliminary empirical results show that this method can approximate key statistical and structural properties of scarce cybersecurity datasets. This suggests that synthetic data may help mitigate data scarcity, and thereby be a valid support for research focusing on the development of data-driven policies in the cybersecurity sector. Moreover, building on these promising results, future research can go a step further by leveraging the contextual knowledge of LLMs specialised for cyber-security applications, such as Foundation-Sec-8B**.**

References

[1]. T. Khot, H. Trivedi, et al., Decomposed prompting: a modular approach for solving complex tasks, *arXiv preprint*, 2022, arXiv:2210.02406.

[2]. P. Liu, W. Yuan, et al., Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing, *arXiv preprint*, 2021, arXiv:2107.13586.

[3]. A. Jolicoeur-Martineau, et al., Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees, in *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS'24)*, 2024, pp. 1288-1296.

[4]. S. Onishi, S. Meguro, Rethinking data augmentation for tabular data in deep learning, *arXiv preprint*, 2023, arXiv:2305.10308.

[5]. R. Sauber-Cole, T. M. Khoshgoftaar, The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey, *Journal of Big Data*, Vol. 9, Issue 1, 2022, 62.

[6]. X. Fang, W. Xu, et al., Large language models (LLMs) on tabular data: prediction, generation, and understanding – a survey, *arXiv preprint*, 2024, arXiv:2402.17944.

[7]. T. Liu, Z. Qian, et al., Goggle: generative modelling for tabular data by learning relational structure, in *Proceedings of The Eleventh International Conference on Learning Representations (ICLR'23)*, 2023.

[8]. Y. Sui, T. Wu, et al., Self-supervised representation learning from random data projectors, *arXiv preprint*, 2023, arXiv:2310.07756.

[9]. J. Wei, X. Wang, et al., Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems*, Vol. 35, 2022, pp. 24824-24837.

[10]. V. Borisov, K. Seßler, et al., Language models are realistic tabular data generators, *arXiv preprint*, 2022, arXiv:2210.06280.

[11]. A. V. Solatorio, O. Dupriez, Realtabformer: generating realistic relational and tabular data using transformers, *arXiv preprint*, 2023 arXiv:2302.02041.

[12]. T. Zhang, S. Wang, et al., Generative table pre-training empowers models for tabular prediction, *arXiv preprint*, 2023, arXiv:2305.09696.

[13]. Z. Zhao, R. Birke, L. Y. Chen, Tabula: harnessing language models for tabular data synthesis, in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'25)*, 2025, pp. 247-259.

[14]. Unsloth AI, unsloth/Llama-3.2-1B, Computer software, Hugging Face, 2025. https://huggingface.co/unsloth/Llama-3.2-1B

[15]. N. Park, M. Mohammadi, et al., Data synthesis based on generative adversarial networks, *Proceedings of the VLDB Endowment*, Vol. 11, Issue 10, 2018, pp. 1071-1083.

(015)

Unforgetting Educational Surveillance: Reimagining AI as a Tool for Justice and Pedagogical Liberation

G. Parker
University Canada West, Leadership and People Management Department, 626 W Pender St #100, Vancouver, BC V6B 1V9
Tel.: +1 236.259.1821
E-mail: gifty.parker@ucanwest.ca

Summary: This study critically investigates how institutional policies and pedagogical practices in Canadian higher education construct, regulate, and resist the integration of generative AI. Drawing on Critical Discourse Analysis [1], and narrative inquiry, the paper analyzes policy documents from select universities and centers the lived experiences of graduate students – particularly international and multilingual learners. The research foregrounds how AI governance intersects with academic surveillance, linguistic hierarchies, and digital equity. Through participatory practices such as AI journaling and policy prototyping, the study advances ethical, justice-oriented approaches that challenge exclusionary norms embedded in dominant AI discourses. Findings highlight that current frameworks often reproduce epistemic inequities under the guise of academic integrity. By interrogating the sociocultural and ecological dimensions of AI infrastructure, this work contributes to debates on ethical data governance, explainability, and student agency. The study proposes a reimagining of AI policy as participatory, inclusive, and aligned with decolonial and critical pedagogical principles.

Keywords: Generative AI, Critical pedagogy, Epistemic justice, Curriculum studies, Educational policy, AI Governance in education, Data ethics in higher education.

1. Introduction

This paper investigates how generative AI is being governed in higher education through restrictive policy frameworks that mirror long-standing traditions of educational surveillance. As universities scramble to address AI's implications for student work, the resulting regulations often position AI as a threat to authenticity and learning, reinforcing assumptions about authorship and academic integrity that reflect deeper anxieties about automation, labor, and knowledge production. Drawing on Critical Discourse Analysis [2, 3], and narrative inquiry, the study analyzes policy texts from multiple institutions alongside reflective accounts from faculty navigating new AI rules in their classrooms.

The findings suggest that dominant institutional responses to generative AI, echoing insights from [4] and [5], rely on control-based models of teaching, in which educators and students are positioned as potential violators rather than co-creators. Selwyn cautions that the deployment of AI in education often reinforces hierarchical structures and managerial logics, sidelining democratic and relational pedagogies. Watters, tracing the history of teaching machines, similarly illustrates how educational technologies have historically been used to regulate behavior and enforce standardization rather than support emancipatory learning. These frameworks are deeply entangled with the broader datafication[1] of education, as described by [6], where automated decision-making and surveillance logics increasingly shape pedagogical practices. By foregrounding narratives from the classroom, this study draws attention to how such policies marginalize alternative epistemologies, reinforcing the kinds of racialized and classed exclusions highlighted by [7-9] in their critiques of algorithmic injustice and institutional gatekeeping.

Rather than proposing AI as an uncritical solution or a pedagogical threat, the paper frames it as a contested site of meaning-making, one that can either reproduce dominant power structures or serve as a catalyst for critical, inclusive, and participatory forms of knowledge. In response, the paper advocates for data-informed, justice-oriented pedagogical frameworks that foreground epistemic diversity and critical data literacies, challenging the binary of "AI misuse" versus "proper use" that dominates policy discourse. Ultimately, it calls for a reimagination of AI's role in education – not as a system to be policed, but as a tool to rethink authorship, agency, and learning in data-driven times.

As [10] provocatively asserts, constructing radically different futures necessitates redistributing power and agency to those historically marginalized and excluded from institutional educational systems. This assertion fundamentally unsettles the prevailing logics shaping higher education's reaction to generative artificial intelligence (AI). Across numerous institutions, including those that position themselves as progressive – AI integration is not

[1] Datafication refers to the transformation of social action into quantified data for analysis and decision-making.

embraced as a transformative pedagogical opportunity but is instead rigidly contained within restrictive policy frameworks, quantitative quotas, and intensified surveillance mechanisms. These strategies are not neutral but rather perpetuate a legacy of control and disciplinary governance within education. For instance, some universities impose a 20 % cap on AI-generated content in student submissions. Yet, the questions of who arbitrates this threshold and why AI use must be relegated to such margins remain unaddressed. This paper critiques these arbitrary limits as symptomatic of broader institutional imperatives designed to police epistemic boundaries and uphold entrenched hierarchies. Such practices betray an educational system grappling with profound epistemological ruptures: What does learn, authorship, and knowledge production mean when the distinction between human and machine agency is destabilized? Generative AI platforms like ChatGPT and DALL·E do not merely assist students – they challenge deeply ingrained assumptions of originality, voice, and assessment that are rooted in Western academic traditions and sustained through systematic exclusion [11].

Dominant academic norms have long marginalized students for whom English functions as an additional or second language, as well as those who engage in culturally situated and non-normative epistemologies. The deployment of AI detection technologies – purportedly to identify "non-authentic" or machine-generated writing, frequently misclassifies multilingual and unconventional linguistic expressions as fraudulent. This misrecognition is far from a benign technical flaw; it constitutes a form of linguistic erasure that disproportionately punishes those already navigating colonial and exclusionary educational structures. What is exposed here is not AI as an inherent threat to academic integrity but rather the institutional apparatus that wields AI tools as instruments of gatekeeping, policing who is authorized to speak, how they are allowed to speak, and which modes of knowledge are legitimated. Such practices reinforce colonial legacies[1] of epistemic violence and perpetuate exclusion through technological means.

2. Research Questions and Participant Justification

This study's research questions are critically designed to interrogate the ways institutional power, knowledge production, and identity are reconfigured in response to the integration of generative AI within higher education. By centering policy discourse and the lived educational experiences of students and educators, these questions illuminate the mechanisms through which AI governance simultaneously reinscribes and challenges entrenched epistemic hierarchies, particularly those shaped by linguistic, cultural, racial, and migratory inequities.

The first question critically examines institutional AI policies as discursive sites where power relations and ideological constructs converge to regulate academic integrity and authorship. Drawing on the foundational framework of Critical Discourse Analysis articulated by [1], these policies are understood not as neutral administrative documents but as strategic texts that enact surveillance, control, and normative discipline. This surveillance disproportionately privileges dominant Western, monolingual conceptions of knowledge, authority, and academic legitimacy [6], thereby perpetuating exclusionary academic norms as highlighted in scholarship on the datafication of education [6, 4]. Through this lens, the study reveals how generative AI is embedded within broader institutional regimes that marginalize alternative epistemologies, reinforcing patterns of epistemic injustice.

The second and third questions engage directly with the complex lived realities of educators and students – particularly those who identify as international, multilingual, and from historically marginalized backgrounds. The critical pedagogical tradition underscores the necessity of centering marginalized voices to challenge and dismantle hegemonic academic cultures [12, 13]. Many international students, who often settle domestically after migration, navigate complex identity negotiations and epistemic tensions that frequently clash with prevailing Western academic norms privileging specific modes of authorship and linguistic expression [14, 15]. By exploring their experiences with AI tools in educational settings, this inquiry highlights how institutional policies translate into practice, affecting student agency, trust, and equitable access. An intersectional framework further informs this analysis, attending to how race, language, and migration status intersect to shape layered educational inequities [16, 7].

The fourth research question foregrounds participatory and justice-oriented pedagogies as essential interventions to reimagine AI governance within academia. Informed by decolonial and feminist epistemologies, this inquiry advances the imperative that students must be active co-creators of knowledge and governance frameworks to resist the disciplinary and surveillance logics embedded within AI systems [17, 18]. Participatory methods such as AI journaling and policy prototyping foster critical reflexivity and collective agency, enabling disruptions to punitive, top-down governance models and promoting epistemic

[1] Colonial legacies of epistemic violence" refers to historical and ongoing practices through which dominant knowledge systems marginalize, silence, or devalue the ways of knowing of colonized or minoritized groups [19, 20]. In the context of AI-mediated assessment, these legacies are reflected in technologies and policies that disproportionately disadvantage multilingual or culturally diverse students.

sovereignty[1] [19, 20], here understood as students' capacity to exercise control over knowledge construction and evaluation. These approaches are especially vital for recentering the voices of students from diverse linguistic and cultural backgrounds who have been historically marginalized within dominant academic paradigms.

Lastly, this study situates AI governance within broader ecological and sociocultural frameworks, responding to urgent calls for critical environmental pedagogy that foregrounds the interconnections between social justice and environmental sustainability [4, 21]. Drawing on [21]'s emphasis on decolonial and place-based approaches, this holistic perspective acknowledges that AI's material infrastructures – such as data centers, energy consumption, and algorithmic systems are not abstract or neutral but are deeply entangled with systemic inequalities, including colonial legacies, environmental degradation, and marginalized communities' dispossession. Consequently, ethical AI governance must engage intersectional and justice-oriented responses that attend simultaneously to the sociocultural and ecological dimensions of technological development. This approach challenges reductive, purely technical solutions and advocates for integrated frameworks that prioritize both ecological sustainability and epistemic equity.

The deliberate focus on graduate students and educators in Canadian universities – particularly those who are international and multilingual but have settled domestically – reflects the complex intersections of global mobility, linguistic diversity, and institutional power structures. This participant group provides essential insights into how AI policies and pedagogical practices are experienced amidst conflicting academic expectations and cultural norms, offering fertile ground for co-constructing equitable and justice-centered AI governance models [26, 14]. These research questions and participant choices embody a critical commitment to unveiling and challenging the socio-political dimensions of generative AI integration in higher education. They advance a justice-centered research agenda that not only critiques dominant power structures but also imagines transformative pedagogical futures grounded in equity, participation, and decolonial praxis, contributing meaningfully to ongoing debates on privacy, fairness, ethical accountability, and participatory governance in the evolving landscape of AI and big data in education.

3. Methodology and Reflexivity

This study employs a hybrid methodological framework integrating Critical Discourse Analysis (CDA) with narrative inquiry, explicitly grounded in researcher positionality. Institutional AI policies and pedagogical practices are analyzed as discursive texts – dynamic sites where power relations, language, and ideology intersect to construct prevailing definitions of learning, authorship, and academic legitimacy. Drawing on [1] and [3] critical discourse frameworks, this approach reveals how such texts covertly encode mechanisms of control, exclusion, and epistemic policing[2]. Complementing this, I foreground my lived experience as an educator navigating the complexities and contradictions inherent in AI integration within higher education. Informed by feminist, decolonial, and poststructuralist epistemologies, this methodological stance reframes narrative inquiry as a rigorous form of critical knowledge production rather than anecdotal reflection. To enhance analytic depth and writing clarity, generative AI tools (specifically ChatGPT) were employed selectively during initial idea generation and iterative text refinement. These AI-generated outputs were critically evaluated and carefully integrated to maintain alignment with the study's epistemological commitments and reflexive approach. This framework thus provides a lens to interrogate how entrenched histories of surveillance and linguistic marginalization are rearticulated in AI-related institutional policies while simultaneously envisioning alternative pedagogical futures rooted in justice and equity.

3.1. Data Sources and Sampling

This study investigates how institutional AI policies and pedagogical practices shape the integration of generative AI in higher education. The primary data corpus includes policy documents from seven universities across North America and Europe, selected for their accessibility, relevance, and diversity of approaches to AI in teaching and assessment. Complementary pedagogical materials such as course guidelines, assignment rubrics, and faculty communications will be examined to understand how policies are interpreted and operationalized within academic contexts. Document selection prioritizes those published or updated within the past two years, reflecting recent shifts in generative AI capabilities and institutional responses. At the time of writing,

[1] The term 'epistemic sovereignty' originates from decolonial scholarship [19, 20], where it refers to the control Indigenous peoples have over their knowledge systems. In this paper, it is adapted to describe students' capacity to exercise agency and critical control over knowledge in AI-mediated learning environments.

[2] The term "epistemic policing" is used here as an original analytic concept to describe how institutions or dominant knowledge systems regulate who is recognized as a legitimate knower, what counts as valid knowledge, and which forms of expression are authorized. In the context of AI and higher education, this includes policies, assessments, or technologies that disproportionately constrain or marginalize students with non-normative linguistic, cultural, or epistemic practices

initial sampling is underway. Data are being sourced from publicly accessible institutional repositories and websites, supplemented by direct communication with faculty and administrative staff to request internal or unpublished materials. While some limitations may arise due to institutional transparency, the corpus aims to provide a robust cross-section of current policy landscapes. Ethical protocols guide data access, consent, and confidentiality in accordance with institutional research standards.

3.2. Analytical Framework and Procedures

The study employs a hybrid analytical approach that integrates Critical Discourse Analysis (CDA) and Narrative Inquiry, situating the work within a broader decolonial and critical pedagogical tradition. Institutional documents were examined using CDA to surface the latent ideologies, power dynamics, and normative assumptions embedded in policy language. This method is particularly suited for unpacking how AI governance reproduces dominant academic values – such as Western, monolingual, and individualistic conceptions of knowledge, authorship, and surveillance under the guise of academic integrity. Complementing this, Narrative Inquiry was used to explore the lived pedagogical tensions and contradictions encountered in real-time AI integration. Researcher reflections as an active member of an Academic Integrity Committee served as a site of situated knowledge production. These reflections illuminated how institutional policies play out in practice, particularly in cases of improper AI use by students, and how these tensions reveal both the necessity of AI regulation and the opportunity for transformative pedagogical design. Data coding and thematic development are being conducted using qualitative software tools such as NVivo, allowing for iterative, reflexive coding across both institutional documents and narrative data. This methodological integration supports a deep understanding of how policy and pedagogy intersect, and how AI governance can move toward more just, inclusive, and pedagogically sound approaches.

3.3. Participatory Practices in Progress

To deepen the critical and praxis-oriented dimensions of the study, participatory initiatives are currently in early development. These emerging practices seek to engage students not merely as research subjects but as co-constructors of knowledge, aligning with decolonial and justice-centered research ethics. These include: (1) reflective AI journaling, where graduate students engage in structured journaling exercises that document their interactions with AI tools, emphasizing metacognitive awareness and uncovering hidden power dynamics (e.g., reliance, trust, error interpretation). This fosters ethical discernment and agency in AI use; (2) peer-review dialogues, where graduate students participate in dialogic critique exercises collaboratively evaluating human- and AI-generated texts. These sessions challenge dominant narratives of authorship and assess how authority is constructed and contested in academic writing; and (3) participatory policy prototyping, where graduate student groups co-design draft AI usage policies, drawing on their linguistic, cultural, and educational experiences. These co-created policies offer alternatives to punitive models, promoting equity-based governance frameworks that recognize the diversity of international student experiences.

These participatory practices are being designed with particular attention to international student populations and their diverse linguistic and epistemic backgrounds. They aim to challenge hegemonic academic standards and promote a more inclusive AI governance paradigm. Together, these methodologies reflect critical pedagogical commitments – uncovering institutional power in AI policy language, foregrounding the lived realities of educational actors navigating generative AI, and building toward participatory, justice-oriented alternatives that resist reductive, punitive approaches to academic integrity. Data coding and thematic development were conducted using qualitative software tools to support systematic and reflexive engagement. The iterative analysis process drew connections across institutional discourse, student practices, and researcher positionality, allowing for a multi-layered understanding of how generative AI is framed, governed, and negotiated in higher education. By combining institutional critique with participatory pedagogies, this methodology not only analyzes existing systems but also contributes to imagining more just, inclusive, and dialogic futures for AI governance in global academic contexts.

4. Critical Examination of AI Policies and Pedagogy

Estes's assertion that "there is no separation between past and present, meaning that an alternative future is also determined by our understanding of our past" offers a crucial lens through which to interrogate the integration of generative AI in education [11]. Contemporary institutional responses largely characterized by detection, punitive policy enforcement, and restrictive frameworks echo historic modalities of educational surveillance. These range from invasive proctoring technologies to linguistic gatekeeping and racially coded behavioral controls, reproducing longstanding inequities under the guise of academic integrity. Far from advancing transformative academic integrity, these surveillance measures entrench exclusionary practices and deepen mistrust, disproportionately impacting marginalized groups such as neurodivergent students and those who use non-standard English.

As Estes argues [11], compels a rigorous interrogation of how these enduring structures of educational control persist and evolve within current

AI debates. The widespread reliance on detection tools, punitive disciplinary policies, and narrowly defined academic honesty protocols mirrors earlier practices of surveillance that disproportionately impact marginalized groups, including neurodivergent students and non-standard English speakers. Classroom management techniques, too, remain steeped in racialized disciplinary logics. Despite their technological veneer, these systems sustain what Paulo Freire termed a "pedagogy of suspicion" – a critical stance that challenges power by scrutinizing hidden ideologies and inequities in educational practices [13]. This pedagogy reveals how students are presumed untrustworthy, and learning is policed rather than nurtured.

Drawing on Krawec's concept of unforgetting, this paper calls for a critical remembering of educational histories marked by surveillance, exclusion, and epistemic violence [23]. Unforgetting here is not a passive recollection; it is a political and pedagogical practice centered on relationality, historical accountability, and an active refusal of colonial erasure. In the context of AI governance in universities – particularly those serving diverse international student populations – this means interrogating how dominant policies reproduce hegemonic academic standards that privilege Western, monolingual, and individualistic ideas of authorship, knowledge, and language. Such standards disproportionately marginalize students who write in non-standard English, draw from alternative epistemologies, or come from collectivist educational traditions.

Rather than supporting learning, punitive AI detection measures replicate patterns of linguistic and racial surveillance, cloaked in discourses of academic integrity. Unforgetting demands that institutions refuse to normalize these inequities and instead reimagine AI governance as part of a justice-oriented pedagogy – one that embraces multimodality, ethical co-creation, and knowledge sovereignty. This perspective is deepened by Gloria Anzaldúa's Borderlands/La Frontera, which critically examines how colonial and linguistic borders shape identity, knowledge, and power [24]. Anzaldúa's work highlights the lived experience of navigating multiple languages and cultures, challenging hegemonic, monolingual academic norms that AI governance often reinforces. Her insights call for educational policies that validate multilingualism and hybrid epistemologies rather than policing conformity. Additionally, the political urgency emphasized by Tuck and Yang [19, 20] in Decolonization is Not a Metaphor reinforces that addressing AI governance requires more than symbolic reform [20]. They argue that genuine decolonization must confront and dismantle entrenched colonial power structures, rather than merely repurpose them. This demands that universities move beyond superficial diversity initiatives and critically reimagine AI's role in perpetuating or resisting systemic inequities.

4.1. AI-integrated Assessment in Practice

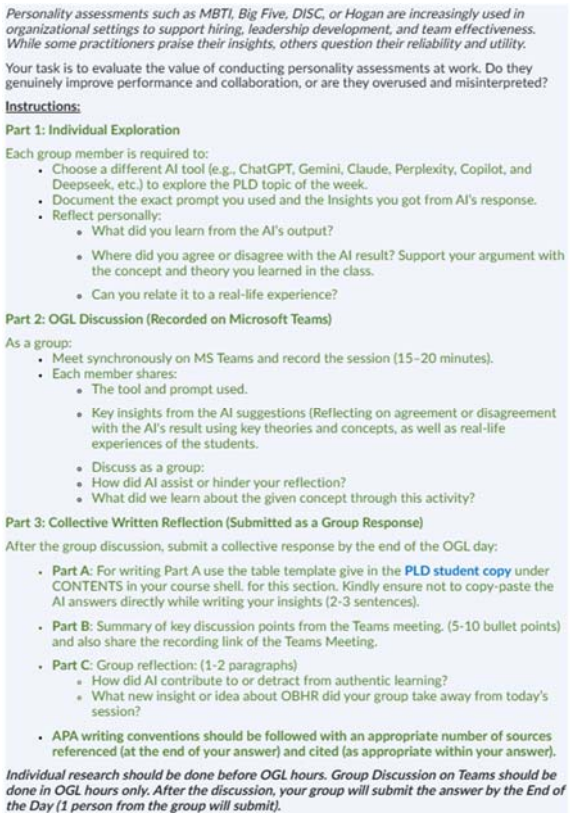
While empirical research on student engagement with generative AI is growing [32-34], this study draws on a different form of empirical evidence: policy documents, course guidelines, assignment rubrics, and faculty communications (see Fig. 1). This document-centered analysis complements broader empirical studies that focus on student experiences, providing a lens to interrogate institutional power, epistemic authority, and pedagogical governance. A close reading of the AI-integrated personality assessment assignment exposes the tension between pedagogical innovation and control. Students are invited to explore multiple AI tools and reflect critically, however, the instructions simultaneously enforce compliance: "Kindly ensure not to copy-paste the AI answers directly." From a CDA perspective, such directives position students as objects of governance, encoding hierarchical power relations and framing AI not as a co-creative partner but as a monitored instrument. Group structures, submission protocols, and reflective prompts further mediate collaboration through oversight, subtly shaping both participation and epistemic legitimacy.

Still within these constraints lie latent possibilities for critical engagement. Reflection and collaborative synthesis offer sites where students might negotiate meaning, contest institutional assumptions, and exercise epistemic agency[1]. The assignment thus emerges as a site of discursive tension: simultaneously a mechanism of control and a potential avenue for emancipation. This prompts a critical question: how might institutional discourses be rearticulated to reposition AI from a compliance tool to a catalyst for justice-oriented pedagogy, equitable knowledge creation, and student-led experimentation? By foregrounding these document-based insights, the study begins to bridge the gap between theoretical critique and practical intervention. The CDA approach uncovers subtle mechanisms through which language and policy mediate learning experiences, providing a foundation for proposing assessment strategies that are both justice-oriented and pedagogically liberatory. While further data collection and participatory insights would deepen the empirical grounding, these preliminary analyses offer a tangible illustration of the tensions and possibilities inherent in ethically and critically integrating AI within higher education.

Extending from this document-centered insight, institutional alignment emerges as a key consideration. Inconsistent expectations across courses, even within

[1] Epistemic agency refers to the learner's capacity to take responsibility for, and actively shape, the processes of knowledge building and validation [35, 36].

the same program, create discursive conditions that constrain student agency and produce epistemic uncertainty, particularly for multilingual or international learners. From a CDA lens, such inconsistencies are not neutral; they actively construct hierarchies of authority, privileging institutional interpretation over student reasoning. Participatory governance frameworks offer a potential corrective: by involving students and faculty in co-creating AI-use policies, institutions can clarify expectations while preserving spaces for critical experimentation. In this approach, reflection, documentation, and collaborative synthesis are both pedagogical strategies and mechanisms through which students negotiate meaning and exercise agency within institutional structures.



Personality assessments such as MBTI, Big Five, DISC, or Hogan are increasingly used in organizational settings to support hiring, leadership development, and team effectiveness. While some practitioners praise their insights, others question their reliability and utility.

Your task is to evaluate the value of conducting personality assessments at work. Do they genuinely improve performance and collaboration, or are they overused and misinterpreted?

Instructions:

Part 1: Individual Exploration

Each group member is required to:

- Choose a different AI tool (e.g., ChatGPT, Gemini, Claude, Perplexity, Copilot, and Deepseek, etc.) to explore the PLD topic of the week.
- Document the exact prompt you used and the Insights you got from AI's response.
- Reflect personally:
  - What did you learn from the AI's output?
  - Where did you agree or disagree with the AI result? Support your argument with the concept and theory you learned in the class.
  - Can you relate it to a real-life experience?

Part 2: OGL Discussion (Recorded on Microsoft Teams)

As a group:

- Meet synchronously on MS Teams and record the session (15–20 minutes).
- Each member shares:
  - The tool and prompt used.
  - Key insights from the AI suggestions (Reflecting on agreement or disagreement with the AI's result using key theories and concepts, as well as real-life experiences of the students.
  - Discuss as a group:
  - How did AI assist or hinder your reflection?
  - What did we learn about the given concept through this activity?

Part 3: Collective Written Reflection (Submitted as a Group Response)

After the group discussion, submit a collective response by the end of the OGL day:

- **Part A**: For writing Part A use the table template give in the **PLD student copy** under CONTENTS in your course shell. for this section. Kindly ensure not to copy-paste the AI answers directly while writing your insights (2-3 sentences).
- **Part B**: Summary of key discussion points from the Teams meeting. (5-10 bullet points) and also share the recording link of the Teams Meeting.
- **Part C**: Group reflection: (1-2 paragraphs)
  - How did AI contribute to or detract from authentic learning?
  - What new insight or idea about OBHR did your group take away from today's session?
- **APA writing conventions should be followed with an appropriate number of sources referenced (at the end of your answer) and cited (as appropriate within your answer).**

Individual research should be done before OGL hours. Group Discussion on Teams should be done in OGL hours only. After the discussion, your group will submit the answer by the End of the Day (1 person from the group will submit).

Fig. 1. Example of AI-integrated personality assessment assignment in a higher education course shell. Note: Using Critical Discourse Analysis, this assignment reflects tensions between pedagogical innovation and institutional control. Language emphasizing documentation, reflection, and prohibition of copy-pasting positions students as subjects of oversight while simultaneously encouraging engagement with AI tools. The structured individual and group tasks reveal underlying power relations and epistemic authority, highlighting opportunities to reframe AI use as a co-creative, justice-oriented learning practice. Source: Anonymized course materials.

Moreover, the discursive framing of AI across programs must be coherent and transparent. Policies, rubrics, and directives should consistently communicate rights and responsibilities while integrating reflection and participatory feedback. CDA highlights that the language used in these texts mediates power and shapes perceptions of legitimacy; transparent, participatory approaches help expose and reconfigure these dynamics. By focusing on critical reflection, ethical reasoning, and student-led documentation of AI engagement, governance structures can transform AI from a monitored tool into a co-creative partner in knowledge production. Ultimately, a justice-oriented perspective extends to broader epistemic and ethical considerations. Decisions about AI use influence whose intellectual contributions are recognized and whose labor is rendered invisible. CDA underscores that knowledge production is always situated within power-laden discourses. Integrating structured reflection, participatory policy-making, and institution-wide guidance allows AI to scaffold equitable, critically engaged learning rather than serve as an instrument of surveillance. In this way, tensions between oversight and innovation become productive sites for critical engagement, positioning AI as a catalyst for emancipatory pedagogy and a vehicle for student agency and epistemic justice.

5. Critical Tensions and Contradictions in AI Governance

Institutional governance of AI in graduate education remains inconsistent, producing a patchwork of rules and surveillance mechanisms that students must navigate. Some courses encourage AI as a reflective, co-creative tool, while others rely on automated detection systems, rigid prohibitions, or arbitrary thresholds (e.g., flagging submissions with less than 20 % AI-generated content). From a CDA perspective [1, 3], these divergent practices are not neutral; they actively construct hierarchies of knowledge and authority, shaping who can claim epistemic legitimacy. International students, already negotiating linguistic and cultural marginality [27, 28], are disproportionately affected, as they must interpret conflicting institutional rules while avoiding disciplinary sanction.

Policing AI, even with ostensibly minor thresholds, enacts a subtle but powerful disciplinary logic. The emphasis on compliance – articulated in directives like "do not copy-paste AI answers" positions students as subjects to be monitored rather than active knowledge producers. CDA highlights how such institutional language discursively enforces control, transforming assignments from sites of inquiry into instruments of governance. Classes without AI oversight, by contrast, may permit exploration but also leave students without clear epistemic guidance, risking confusion about what counts as legitimate knowledge. In both cases, the institutional framing privileges surveillance over critical engagement, raising the question: is the academy more invested in policing behavior than fostering reflective, equitable learning? Moreover, the focus on detection and restriction obscures AI's potential as a pedagogical tool. Generative AI, when

scaffolded appropriately, can extend coverage, support reasoning, and provide alternative modes of understanding – resources particularly valuable for students unfamiliar with dominant academic norms. Yet institutional discourses that foreground risk, rather than opportunity, effectively *"dumb down"* AI, constraining its capacity to enhance learning. CDA exposes this tension: the technology is not inherently disempowering, but institutional practices render it so. Students internalize compliance as epistemic virtue, rather than critical exploration, reinforcing inequities and curtailing agency.

The critical implications are stark. Should graduate programs maintain fragmented, class-specific approaches, or establish coherent, program-wide AI policies that clearly define both rights and responsibilities? Does the act of policing AI genuinely enhance learning, or does it perpetuate systemic inequities and anxiety, especially among international students? How might institutions reframe AI not as a threat to authorship, but as a conduit for reflective, co-creative, and justice-oriented pedagogy? A truly critical approach would disrupt the logic of surveillance, redistribute epistemic authority, and reconceptualize AI as a scaffold for equitable learning. Without such a reframing, AI governance risks reinforcing the very hierarchies it purports to manage, privileging institutional control over student agency, and compliance over critical inquiry.

5.1. Surveillance, Contradiction, and the Politics of AI-integrated Assessment

The coexistence of divergent practices within the same MBA program – one assignment encouraging students to experiment with multiple AI tools, another submission flagged at 95 % AI-generated, alongside a different assignment flagged at 44 % (As shown in Fig. 2 and Fig. 3) – illustrates the fractured and often contradictory discourses that govern AI in higher education. In one space, AI is constructed as an instrument of pedagogical innovation: students are asked to explore, reflect, and collaborate using tools like ChatGPT, Gemini, or Claude. In another, AI is framed as a threat to academic integrity, with algorithmic detection systems legitimating punitive responses. Turnitin's AI writing assessment comes with a disclaimer (see Fig. 2 and Fig. 3) noting that results may be inaccurate and should not be used as the sole basis for disciplinary action. From a CDA perspective, this language is telling: it signals institutional caution while still asserting algorithmic authority, positioning students under surveillance even as the tool admits uncertainty.

For international and multilingual students, these contradictions take on heightened significance. Assignments that encourage experimentation with AI may offer new pathways for support – scaffolding language learning, expanding access to resources, or providing avenues for self-directed exploration [27]. Yet the threat of being flagged by AI-detection tools places these same students under heightened scrutiny. As [29] and [30] argue, language and authorship in academic contexts are never neutral; they are structured by racialized and colonial assumptions about what counts as "authentic" knowledge. AI detection technologies extend these assumptions by treating writing that deviates from standardized linguistic norms as suspect, thereby reinscribing epistemic hierarchies under the guise of technological neutrality.

This contradiction raises a central question: whose knowledge is privileged when AI use is policed, and whose knowledge is erased when it is celebrated? When AI-detection algorithms flag a submission at 95 %, does this erase the student's intellectual labor of prompt design, synthesis, and reflection? Conversely, when AI use is sanctioned in another assignment, is it truly emancipatory, or does it remain bounded by institutional directives that reduce students to compliant users rather than co-creators of knowledge? Here, CDA allows us to see that the question is not simply whether AI use should be permitted or prohibited. Rather, it is about how institutional discourses frame AI in ways that reproduce existing regimes of power. Left unpoliced, AI risks becoming a tool of epistemic outsourcing, where students rely uncritically on machine-generated outputs. Over-policed, it becomes a site of disciplinary surveillance that punishes the very students most likely to benefit from critical engagement with these tools. Both extremes risk undermining the emancipatory potential of AI in education.

A justice-oriented alternative requires reframing the role of AI in assessment. Instead of positioning AI as either a forbidden crutch or a celebrated novelty, institutions could adopt participatory frameworks where students and faculty co-create guidelines for ethical and critical AI use [31]. Reflective AI journals, for instance, could require students to document not just AI outputs but the reasoning behind their use – shifting the focus from product to process, from compliance to critical engagement. Similarly, policy prototyping exercises could invite students to collaboratively reimagine institutional AI policies, positioning students not as passive subjects of surveillance but as active contributors to the governance of their own learning. Ultimately, what is at stake is not only the question of AI in the classroom but the broader struggle over epistemic authority in higher education. If institutions continue to oscillate between uncritical celebration and punitive policing, they risk reproducing a system where students – especially those from marginalized linguistic and cultural backgrounds – remain caught in the paradox of being simultaneously invited and distrusted. To move beyond this impasse, AI must be reimagined not as a compliance tool but as a catalyst for justice-oriented pedagogy: one that foregrounds student agency, values diverse epistemologies, and resists the expansion of educational surveillance into every corner of academic life.

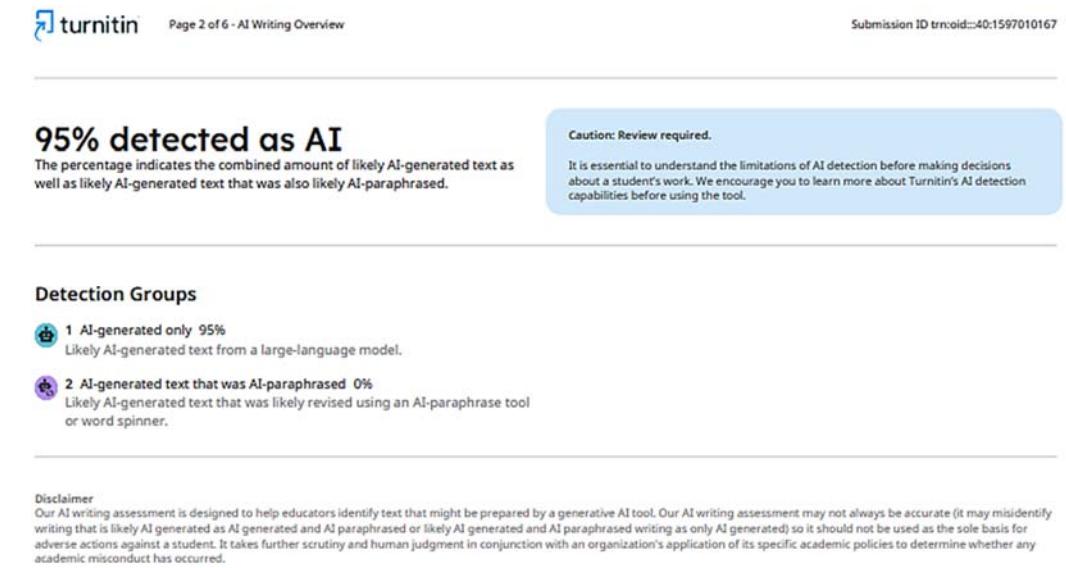


Fig. 2. Turnitin AI-generated content report (95 %). An anonymized example illustrating the detection of high AI-generated content in a student submission.

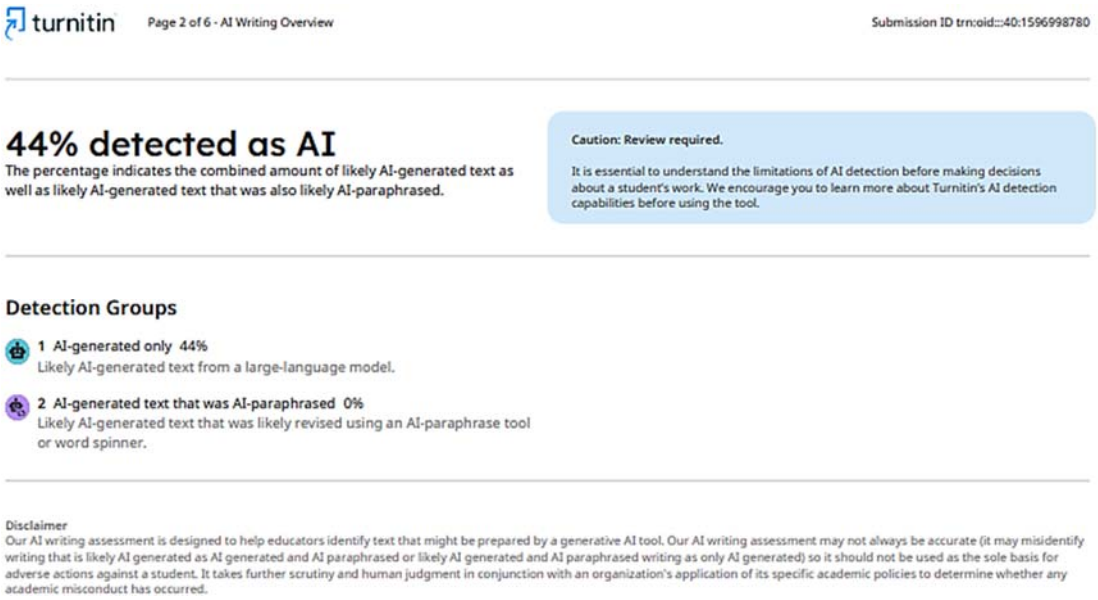


Fig. 3. Turnitin AI-generated content report (44 %). An anonymized example illustrating the detection of moderate AI-generated content in a student submission.

5.2. Toward a Critical Reconceptualization of AI Governance

The institutional imperative to govern AI use in universities often centers on preventing academic dishonesty – a necessity grounded in concerns over the improper deployment of generative AI in student work. Drawing on my experience within the university's Academic Integrity Committee, this paper recognizes the practical need for detection and enforcement mechanisms to uphold academic standards. Cases of undisclosed AI-generated content and superficial engagement with AI tools reveal tangible threats to academic integrity that cannot be ignored.

Yet, this dominant framing – AI as a threat to be policed – functions as a form of disciplinary power that risks reproducing exclusionary logics. Surveillance-based policies disproportionately impact marginalized groups, including international students whose linguistic and epistemological practices diverge from hegemonic academic norms. This punitive orientation often neglects the complex realities of AI's pervasiveness and potential within learning contexts. Critically, governance that privileges control over critical engagement forecloses opportunities for students to develop meaningful literacies around AI. Rather than positioning AI merely as a site of risk, universities must reconceptualize AI governance as a pedagogical intervention – one that cultivates ethical, reflective, and justice-oriented uses of AI. This entails moving beyond reductive detection toward pedagogies that integrate AI literacy, support multimodal

expression, and foster dialogic critique of AI-generated knowledge. Such an approach demands institutional humility and responsiveness: governance must be co-constructed with students, particularly those from marginalized and international backgrounds, acknowledging and validating their linguistic and cultural diversity. As Garcia and Nichols [22] argue, pedagogies of code-switching exemplify how students can critically navigate AI-generated norms without forfeiting their expressive agency.

Ultimately, effective AI governance in higher education cannot be reduced to binary oppositions of policing versus permissiveness. Instead, it must grapple with the contradictions inherent in AI's integration – recognizing its potential both to reinforce inequities and to serve as a tool for critical, emancipatory learning. Only through such a nuanced, justice-centered lens can AI governance transcend surveillance and exclusion to become a catalyst for transformative educational futures.

5.3. Multi-layered Institutional Interactions

Institutional texts, policies, and assignment designs surrounding AI in higher education operate as sites where power, epistemic authority, and compliance are materially and symbolically enacted. Students are simultaneously invited to integrate generative AI into academic writing and constrained by prescriptive language emphasizing surveillance, risk mitigation, and adherence to narrow expectations. For ESL and international students, this dual framing intensifies inequities: navigating linguistic, cultural, and algorithmic norms, they are positioned as both learners and potential violators of institutional standards. Empirical work illustrates that students view AI as a versatile collaborator – supporting brainstorming, drafting, reflection, and revision – yet institutional directives frequently fail to recognize these practices as legitimate forms of learning [34]. Language such as "ensure not to copy-paste AI outputs" (as shown in Fig. 1) discursively positions students as objects of governance, valorizing compliance while marginalizing exploration. This produces a discursive double-bind: students are rhetorically encouraged to engage critically with AI while simultaneously constrained by structures that limit autonomy, experimentation, and epistemic agency.

Fig. 4 operationalizes these dynamics through a multi-layered mapping of institutional and student interactions where power and agency continually interact. At the top layer, institutional and assignment factors encode risk, integrity, and prescriptive AI engagement. Compliance-oriented language and structured tasks shape reflective capacity, while surveillance emphasis discourages epistemic risk-taking, especially for students with lower prior knowledge. Reflection and collaboration components, though embedded in assignments, are circumscribed by broader constraints, producing latent tensions between innovation and control.

Institutional factors (e.g., risk and prescriptive AI policies) exert control, but their effects are mediated by student capitals (cultural, social, epistemic), producing tensions in Layer 3 where constraints and capacities collide. These "cross-layer tensions" both reveal inequities – such as superficial compliance when cultural or epistemic capital is low – and create leverage points for more critical engagement when capital is high. In the second layer, student contextual factors, highlights cultural capital, social capital, and epistemic awareness. Students with higher capital navigate institutional constraints more strategically, engaging critically with AI tools and reflective tasks. Those with lower capital often comply superficially, limiting critical engagement and reinforcing inequities in participation and learning outcomes. Interaction effects form the third layer, revealing how institutional constraints and student capitals converge to shape engagement quality. High-capital students can leverage structured assignments to produce detailed outputs, whereas low-capital students experience constrained epistemic agency. Collaborative structures can either amplify these disparities or mitigate them when scaffolded effectively.

Finally, the flowchart identifies pedagogical opportunities and strategic interventions. Assignments can be scaffolded according to student capital, balancing oversight with co-creative freedom. Peer mentoring and social learning can leverage social capital, while reflective prompts enhance epistemic awareness. These measures redistribute epistemic authority and create pathways for equitable engagement, transforming AI from a monitored instrument into a scaffold for justice-oriented, critically engaged learning. The analysis demonstrates that inequities, tensions, and opportunities in AI-integrated assessment are discursively produced: they emerge from the interaction of institutional discourse, assignment design, and individual student capacities, rather than from the technology itself.

6. Conclusion and Limitations

This paper advocates for a fundamental shift in framing AI integration within higher education – from the limiting question of How much AI use is too much? to the more expansive and generative inquiry: What does learning look like when AI is an ever-present cognitive partner? Such reframing demands pedagogies grounded not in surveillance, suspicion, or fear, but in trust, collaboration, and critical inquiry – foregrounding historically marginalized voices and epistemologies. Drawing on Ngũgĩwa Thiong'o's call to "reclaim the imagination" from colonial logics [25], this paper contends that AI must be repositioned not as a disciplinary tool but as a co-participant in the struggle for pedagogical liberation. Educational AI governance, if uncritically implemented, risks replicating the colonial structures of epistemic exclusion, linguistic policing, and learner mistrust. To resist this, institutions must imagine AI frameworks

that affirm multilingualism, epistemic plurality, and student agency. Crucially, as Selwyn reminds us, the material costs of AI, its environmental extraction, infrastructural demands, and embedded inequalities – must not be treated as peripheral [4]. Sustainability is integral to educational justice. Ethical AI governance must therefore account for both its sociopolitical and ecological dimensions.

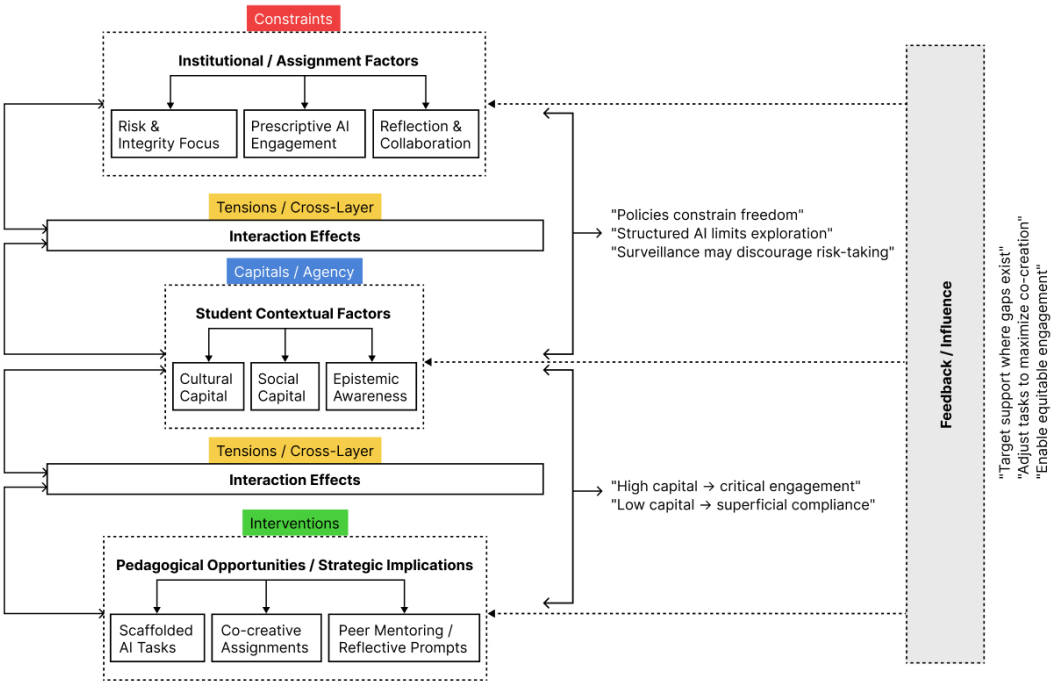


Fig. 4. Multi-Layered Flow of Institutional Constraints, Student Capitals, Interaction Effects, and Pedagogical Opportunities in AI-Integrated Learning. Note: Conceptual work and all ideas represented in this figure are original to Gifty Parker, PhD. The visual diagram was executed by Yaprak Deniz Yurt, PhD purely as a rendering of the conceptual design. The figure demonstrates how institutional policies, student contexts, cross-layer tensions, and pedagogical interventions interact to shape AI engagement, reflection, and equitable outcomes.

Moreover, critical engagement requires unforgetting the rich digital literacies cultivated by Indigenous, Black, and multilingual communities – practices often marginalized in mainstream AI discourse but essential for epistemic sovereignty and cultural survival [7, 8]. Integrating these legacies challenges hegemonic academic norms and opens pedagogical possibilities for ethically scaffolded learning, including "pedagogies of code-switching" [22] that support students in navigating the boundaries between human and AI-generated expression with critical awareness and ethical care. Despite these contributions, this study has limitations. It draws primarily on document analysis and narrative inquiry from a purposive sample of institutions, which may not fully reflect the global diversity of AI practices or institutional responses. While reflexivity helps mitigate interpretive bias, the subjective positioning of the researcher inevitably shapes the analysis. Future research must expand to include broader, direct engagement with students and educators, particularly those from marginalized backgrounds to co-create AI governance models that reflect lived realities and prioritize justice. In closing, this work contributes foundational insights for reconceptualizing AI not as a threat to academic integrity, but as a contested terrain where equity, power, and imagination must be negotiated. Reclaiming AI's potential for liberatory pedagogy demands sustained critical vigilance, participatory research, and ethical commitment. Only then can educational institutions foster AI futures that are not merely technologically advanced, but socially just, ecologically responsible, and pedagogically transformative.

References

[1]. N. Fairclough, R. Wodak, Critical discourse analysis, in Discourse Studies: A Multidisciplinary Introduction, (T. van Dijk, Ed.), Vol. 2, *Sage*, 1997, pp. 258-284.

[2]. T. Van Leeuwen, Discourse and Practice: New Tools for Critical Discourse Analysis, *Oxford University Press*, 2008.

[3]. T. A. van Dijk, Principles of critical discourse analysis, Discourse & Society, Vol. 4, Issue 2, 1993, pp. 249-283.

[4]. N. Selwyn, On the limits of artificial intelligence (AI) in education, *Nordisk Tidsskrift for Pedagogikk & Kritikk*, Vol. 10, Issue 1, 2024.

[5]. A. Watters, Teaching Machines: The History of Personalized Learning, *The MIT Press*, 2021.

[6]. B. Williamson, Big Data in Education: The Digital Future of Learning, Policy and Practice, 1st Ed., *SAGE Publications*, 2017.

[7]. R. Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code, *Polity Press*, 2019.
[8]. S. U. Noble, Algorithms of Oppression: How Search Engines Reinforce Racism, *New York University Press*, 2018.
[9]. T. M. Cottom, Lower Ed: The Troubling Rise of For-Profit Colleges in the New Economy, *The New Press*, 2017.
[10]. P. Bisht, Decolonizing futures: exploring storytelling as a tool for inclusion in foresight, Master's Thesis, *OCAD University*, 2017.
[11]. N. Estes, Our History Is the Future: Standing Rock Versus the Dakota Access Pipeline, and the Long Tradition of Indigenous Resistance, *Verso Books*, 2019.
[12]. b. hooks, Teaching to Transgress: Education as the Practice of Freedom, *Routledge*, 1994.
[13]. P. Freire, Pedagogy of the oppressed, in Toward a Just World Order, Vol. 1, *Routledge*, 2019, pp. 47-54.
[14]. S. Canagarajah, Translingual Practice: Global Englishes and Cosmopolitan Relations, *Routledge*, 2012.
[15]. O. Garcia, L. Wei, Translanguaging: Language, Bilingualism and Education, 1st Ed., *Palgrave Macmillan*, 2014.
[16]. K. Crenshaw, Mapping the margins: intersectionality, identity politics, and violence against women of color, *Stanford Law Review*, Vol. 43, Issue 6, 1991, pp. 1241-1299.
[17]. W. D. Mignolo, The Darker Side of Western Modernity: Global Futures, Decolonial Options, *Duke University Press*, 2011.
[18]. M. Lugones, Heterosexualism and the colonial/modern gender system, *Hypatia*, Vol. 22, Issue 1, 2007, pp. 186-209.
[19]. L. T. Smith, Decolonizing Methodologies: Research and Indigenous Peoples, 3rd Ed., *Zed Books*, 2021.
[20]. E. Tuck, K. W. Yang, Decolonization is not a metaphor, Decolonization: Indigeneity, *Education & Society*, Vol. 1, Issue 1, 2012, pp. 1-40.
[21]. E. Tuck, M. McKenzie, Place in Research: Theory, Methodology, and Methods, *Routledge*, 2015.
[22]. A. E. Garcia Quintana, C. A. Nichols, Code switching and the Hispanic consumer: the effects of acculturation on the language of advertising among Hispanics, *Hispanic Journal of Behavioral Sciences*, Vol. 38, Issue 2, 2016, pp. 222-242.
[23]. P. Krawec, Becoming Kin: An Indigenous Call to Unforgetting the Past and Reimagining Our Future, *Broadleaf Books*, 2022.
[24]. G. Anzaldúa, Borderlands = La Frontera: The New Mestiza, Critical ed. (R. F. Vivancos-Pérez, N. E. Cantú, Eds.), *Aunt Lute Books*, 2021.
[25]. N. wa Thiong'o, Something Torn and New: An African Renaissance, 1st Ed., *BasicCivitas Books*, 2009.
[26]. G. J. S. Dei, Reframing Blackness and Black Solidarities Through Anti-Colonial and Decolonial Prisms, *Springer*, 2017.
[27]. A. S. Canagarajah, Teacher development in a global profession: an autoethnography, *TESOL Quarterly*, Vol. 46, Issue 2, 2012, pp. 258-279.
[28]. D. Dippold, M. Heron, K. Gravett, International students' linguistic transitions into disciplinary studies: a rhizomatic perspective, *Higher Education*, Vol. 83, Issue 3, 2022, pp. 527-545.
[29]. G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, et al., Shifting attention to accuracy can reduce misinformation online, *Nature*, Vol. 592, 2021, pp. 590-595.
[30]. N. Flores, J. Rosa, Undoing appropriateness: raciolinguistic ideologies and language diversity in education, *Harvard Educational Review*, Vol. 85, Issue 2, 2015, pp. 149-171.
[31]. E. Sporrong, C. McGrath, T. Cerratto Pargman, Situating AI in assessment – an exploration of university teachers' valuing practices, *AI & ETHICS*, Vol. 5, Issue 3, 2025, pp. 2381-2394.
[32]. F. Guo, L. Zhang, T. Shi, H. Coates, Whether and when could generative AI improve college student learning engagement?, *Behavioral Sciences*, Vol. 15, Issue 8, 2025, 1011.
[33]. M. Liu, Y. Ren, L. M. Nyagoga, F. Stonier, et al., Future of education in the era of generative artificial intelligence: consensus among Chinese scholars on applications of ChatGPT in schools, *Future in Educational Research*, Vol. 1, Issue 1, 2023, pp. 72-101.
[34]. J. Kim, S. Yu, R. Detrick, N. Li, Exploring students' perspectives on generative AI-assisted academic writing, *Education and Information Technologies*, Vol. 30, Issue 1, 2025, pp. 1265-1300.
[35]. M. Scardamalia, Collective cognitive responsibility for the advancement of knowledge, in Liberal Education in a Knowledge Society (B. Smith, Ed.), *Open Court*, 2002, pp. 67-98.
[36]. C. I. Damşa, P. A. Kirschner, J. E. B. Andriessen, G. Erkens, et al., Shared epistemic agency: an empirical study of an emergent construct, *The Journal of the Learning Sciences*, Vol. 19, Issue 2, 2010, pp. 143-186.

(016)

Enhancing Network Intrusion Detection Using Advanced Meta-learning Ensemble SVMs in Production Cloud Environments

Lokesh Karanam [1] and Hardik Mahant [2]
[1] Independent Researcher, Austin, Texas, USA
[2] Independent Researcher, San Jose, California, USA
E-mail: lokeshkaranam3@gmail.com, hardik.s.mahant@gmail.com

Summary: This paper offers an in-depth study of meta-learning ensemble Support Vector Machine (SVM) models orientated towards resolving the persistent problem of class imbalance in network intrusion detection systems. Modern intrusion detection systems face grave class imbalance problems in terms of detection for sophisticated, relatively scarce attacks like Remote-to-Local (R2L) and User-to-Root (U2R) which constitute less than one percent of network traffic, but wreak disproportionate havoc. To address the challenges posed by the detection of the minority class while keeping the computational cost manageable for real-world applications, we implemented and analyzed eight meta-learning ensemble SVM models: OneVsOne, Pairwise Meta, Enhanced Meta, Balanced Meta, and Focused Minority. Using the NSL- KDD dataset which contains 125973 training samples and 22544 test samples, we were able to perform extensive experiments and our Focused Minority SVM ensemble outperformed other models achieving 77.52 % accuracy, an improvement of 3.5 % over traditional OneVsOne SVMs. With these models, we were able to achieve detection rates ranging from 0.5 % to 10.3 % for R2L and 9 % to 14.9 % for U2R attacks. The study offers new meta- learning frameworks incorporating class imbalance adaptive strategies, weighted binary classifiers, and probabilistic feature enhancement and ensemble voting mechanisms optimized for real-time production deployment, enabling security operations centers to maintain effective intrusion detection with manageable false positive rates in high-traffic enterprise networks.

Keywords: Intrusion detection, Support vector machines, Meta-learning, Ensemble methods, Class imbalance, Cybersecurity.

1. Introduction

The weaponization of artificial intelligence and the growing interconnectedness of digital ecosystems have led to an unprecedented evolution in cyberattacks. With AI-generated phishing emails achieving a 78 % open rate and a 202 % increase in phishing email messages, modern threat actors are using generative AI technologies to craft more convincing phishing campaigns [1]. Ransomware threat landscape has grown dramatically, with attacks rising by 11 % in 2024 to 5414 reported incidents globally [2]. Supply chain attacks have become a particularly destructive attack vector, increasing by 431 % between 2021 and 2023. The post-pandemic digital transformation has further exacerbated this escalating threat environment, with 80 % of organizations worldwide experiencing cloud security breaches and cyberattacks more than doubling in frequency [3]. Financial institutions are especially vulnerable, accounting for 45 % of attacks on critical infrastructure and paying an average of $4.4 million for data breaches.

Traditional intrusion detection systems are seriously flawed, with only 17 % prevention rates against sophisticated ransomware threats and inadequate detection of minority class attacks. Class imbalance problems in network traffic data further complicate the problem. Conventional machine learning techniques, such as standard Support Vector Machines, encounter challenges in defining precise decision boundaries for these sparsely represented attack categories, leading to elevated false negative rates for significant threats. This study addresses significant constraints by proposing and evaluating advanced meta-learning ensemble SVM frameworks specifically designed for minority class detection within real-time production cloud systems [5] that incorporate multiple specialized models utilizing diverse kernels, adaptive class weighting techniques, and probabilistic feature augmentation strategies. This approach establishes a network of specialized classifiers that engage in knowledge exchange via ensemble voting mechanisms, leveraging the collaborative potential of multiple SVM models to identify attack patterns overlooked by traditional single-model methodologies.

2. Literature Review

2.1. Traditional Machine Learning Approaches in Intrusion Detection

Over the past decade, there has been significant progress in the use of machine learning techniques for network intrusion detection [6], with Support Vector Machines being identified as a particularly successful approach for binary and multiclass classification problems. The study by Chowdhury et al [7] demonstrated the effectiveness of SVM-based techniques using randomly generated feature sets, achieving acceptable results on known datasets and highlighting the critical function of feature selection in intrusion detection. However, their approach had limitations when dealing with issues of class imbalance, particularly when it came to determining

which advanced minority class attacks presented the biggest security threats in operational environments.

Scholars have studied a variety of dimensionality reduction techniques to improve classifier efficacy in order to address the fundamental problem of feature selection in intrusion detection [8]. In their study, Alijawarneh et al [9] presented hybrid models that combine Decision Trees, Naive Bayes, and other traditional approaches, demonstrating the effectiveness of ensemble approaches and reducing the number of features from 41 to 8. Despite these developments, traditional approaches continue to encounter difficulties with the large class imbalance in real-world network traffic data, where minority attack types are routinely underrepresented in training datasets. Automated Machine Learning (AutoML) techniques have demonstrated potential for lowering manual overhead while attaining better performance. Gyimah et al. showed that AutoML-driven stacked ensemble models outperformed individual models such as Random Forest and conventional boosting techniques, achieving 90 % accuracy and 89 % F1 score on the NSL-KDD dataset [10].

In intrusion detection systems, the preprocessing stage is acknowledged as a crucial component that affects the system's overall performance. Traditional approaches tend to focus on feature normalization, categorical encoding, and simple sampling, but they do not address the fundamental issues of minority class detection. Recent studies [11] have highlighted the limitations of conventional preprocessing methods when dealing with extremely unbalanced datasets, especially when minority classes such as R2L and U2R attacks require customized strategies to achieve acceptable detection rates.

2.2. Deep Learning and Neural Network Approaches

The advent of deep learning techniques has opened up novel avenues for intrusion detection, with Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks demonstrating encouraging outcomes across a range of cybersecurity applications. The work by Karanam et al [12] introduced a hybrid architecture combining CNN and LSTM, which attained a training accuracy of 99.6 % and a testing accuracy of 89.23 % on the NSL-KDD dataset, thereby illustrating the efficacy of deep learning methodologies in the domain of intrusion detection. Their methodology focused on enhancing computational efficiency through the conversion of feature vectors into matrix representations that are appropriate for convolutional processing, thereby significantly decreasing both parameter count and training duration. Recent methods like TG-LSTM [13] leverages temporal prediction and customized deep learning architectures to improve minority event detection accuracy in imbalanced datasets. Recent studies have demonstrated that GNN-based approaches achieve 98 % accuracy for binary classification and 99.20 % for multi-class classification by modeling network flows as graph structures [14]. GNNs are an emerging paradigm that captures structural relationships in network data. With attention mechanisms that successfully capture temporal dependencies in network traffic patterns, transformer architectures derived from natural language processing have shown remarkable ability in sequence modeling for cybersecurity applications, attaining 96 % accuracy in zero-day threat detection [15].

Nonetheless, deep learning methods [16] encounter numerous significant constraints when implemented in production settings for intrusion detection. The computational complexity associated with training and inference operations frequently surpasses the real-time demands of production systems, where response times under 100 milliseconds are critical for effective threat mitigation. Furthermore, deep learning models generally necessitate a substantial amount of training data to attain peak performance, which may be lacking for minority attack classes in practical situations.

One significant limitation of deep learning approaches in the field of cybersecurity is interpretability. Particularly in crucial circumstances involving possible security incidents, production security systems require transparent decision-making procedures that enable security analysts to understand and validate detection results. This requirement is complicated by the opaque nature of deep neural networks, which makes it difficult for security teams to respond to and trust automated detection results.

2.3. Ensemble Methods and Meta-learning in Cybersecurity

Ensemble learning techniques, which combine multiple specialized models, have gained attention in cybersecurity research for improving detection effectiveness. Traditional ensemble methods, like bagging, boosting, and stacking, have been used to address class imbalance problems, but most focus on homogeneous base learners. Meta-learning, or "learning to learn," is a revolutionary departure from traditional machine learning techniques, allowing systems to adapt their learning strategies based on past experiences. It can address inadequate training data for minority attack classes in intrusion detection. Recent studies [17] have shown improved performance in few-shot learning scenarios where traditional approaches face data scarcity. Combining meta-learning concepts with ensemble approaches can create robust detection systems that learn from a small number of examples while maintaining high accuracy.

2.4. Class Imbalance and Minority Class Detection

Class imbalance in machine learning applications in cybersecurity is a significant issue. Traditional methods like cost sensitive learning frameworks,

undersampling techniques, and SMOTE struggle to address complex, dynamic patterns in intrusion detection environments. SMOTE can increase recall rates but introduces noise and overfitting issues. Cost-sensitive learning approaches distribute higher misclassification costs to minority classes during training, showing promise in cybersecurity. However, finding the best cost matrices is challenging in dynamic environments with varying attack patterns.

2.5. Research Gap and Motivation

Despite extensive research into machine learning methods for intrusion detection, there are still significant challenges in minority class detection. Traditional single-model or homogeneous ensemble approaches fail to leverage the unique advantages of different classifier architectures. Real-time production environments require response times of less than 100 milliseconds for effective threat mitigation. Current research has mainly focused on static benchmark datasets, neglecting the dynamic nature of network environments. To overcome these limitations, meta-learning concepts can be combined with specialized ensemble architectures, but careful planning and evaluation are needed. Advanced meta-learning ensemble SVM frameworks are being developed to address these gaps.

2.6. Traditional SVM and Ensemble Methods

SVMs excel at binary and multi-class classification [18, 7], but challenge in multiclass, imbalanced data. Prior works include a hierarchical SVM ensemble for image classification [19], and hybrid models for network security [9].

Deep learning (CNN-LSTM hybrids) improves detection but incurs high computational cost and lacks explainability [12]. Meta-learning, combining multiple specialized models, is emerging for rare-class detection and incremental adaptation [20]. Recent ensemble SVMs outperform single SVMs for both tabular and structured data tasks [19].

2.7. Class Imbalance Techniques

Imbalanced learning handles rare events by balancing training samples or weighting classes [21]. Methods include SMOTE oversampling, cost-sensitive loss, and ensemble stacking. Recent research optimizes SVM hyperparameters and feature sets hierarchically for minority class gain [19].

3. Methodology

3.1. Dataset Description and Preprocessing

The NSL-KDD [4, 22] dataset, an improved version of the KDD Cup 1999 dataset, was used in an experimental evaluation to overcome the drawbacks of earlier intrusion detection benchmarks. The dataset provides a more balanced depiction of network traffic patterns, eliminating redundant records and dividing each connection record into traffic-based, content-based, and basic features. The dataset comprises 22544 testing samples and 125973 training samples. The dataset's stark class imbalance reflects real network traffic distributions, with malicious activity frequency significantly lower than normal connections. The training dataset includes 45927 Denial of Service (DoS) attacks (36.46 %) and 67343 normal connections (53.46 %). The significant difference between R2L and U2R attack categories presents challenges for traditional machine learning methods and emphasizes the need for customized ensemble approaches [23]. A comprehensive methodology for feature engineering and data standardization was used in the preprocessing pipeline, ensuring consistency between training and testing datasets. StandardScaler normalization was used to standardize features, improving the convergence characteristics of Support Vector Machine-based classifiers. Normal traffic was classified as class 0, DoS attacks as class 1, probe attacks as class 2, R2L attacks as class 3, and U2R attacks as class 4.

3.2. Meta-learning Ensemble Architecture Design

The proposed meta-learning ensemble framework (Fig. 1) integrates several specialized binary classifiers, which are amalgamated through advanced voting mechanisms to tackle the challenges associated with minority class detection. The architecture is composed of three fundamental components: the generation of specialized binary classifiers, the enhancement of probabilistic features, and the fusion of ensemble decisions. This hierarchical methodology facilitates the system’s ability to discern optimal decision boundaries for successive class pairs, while utilizing meta-learned features to enhance overall classification accuracy.

The process of developing binary classifiers (Fig. 2) entails building customized support vector machine models that are trained on various class combinations. Several kernel functions are used in this procedure, and hyperparameter configurations customized for each distinct classification task are optimized. Depending on the complexity of the decision boundary required to differentiate the specific class combination, a binary SVM classifier is constructed using RBF, polynomial, or linear kernels for each successive class pair (i, j). The rationale behind this approach stems from the understanding that different kinds of attacks exhibit distinct feature patterns that might be better captured by specialized classifiers rather than depending on a single multi-class model.

The probabilistic feature enhancement component utilizes the outputs of decision functions from binary classifiers to create supplementary features for the ultimate ensemble classifier. For every binary

classifier developed for the class pair (i, j), probabilistic estimates are derived through the application of the sigmoid function to the decision function values, thereby offering confidence measures regarding class membership predictions. The probabilistic features are integrated with the original feature vectors to form improved representations that encompass both fundamental network statistics and derived classification confidence measures.

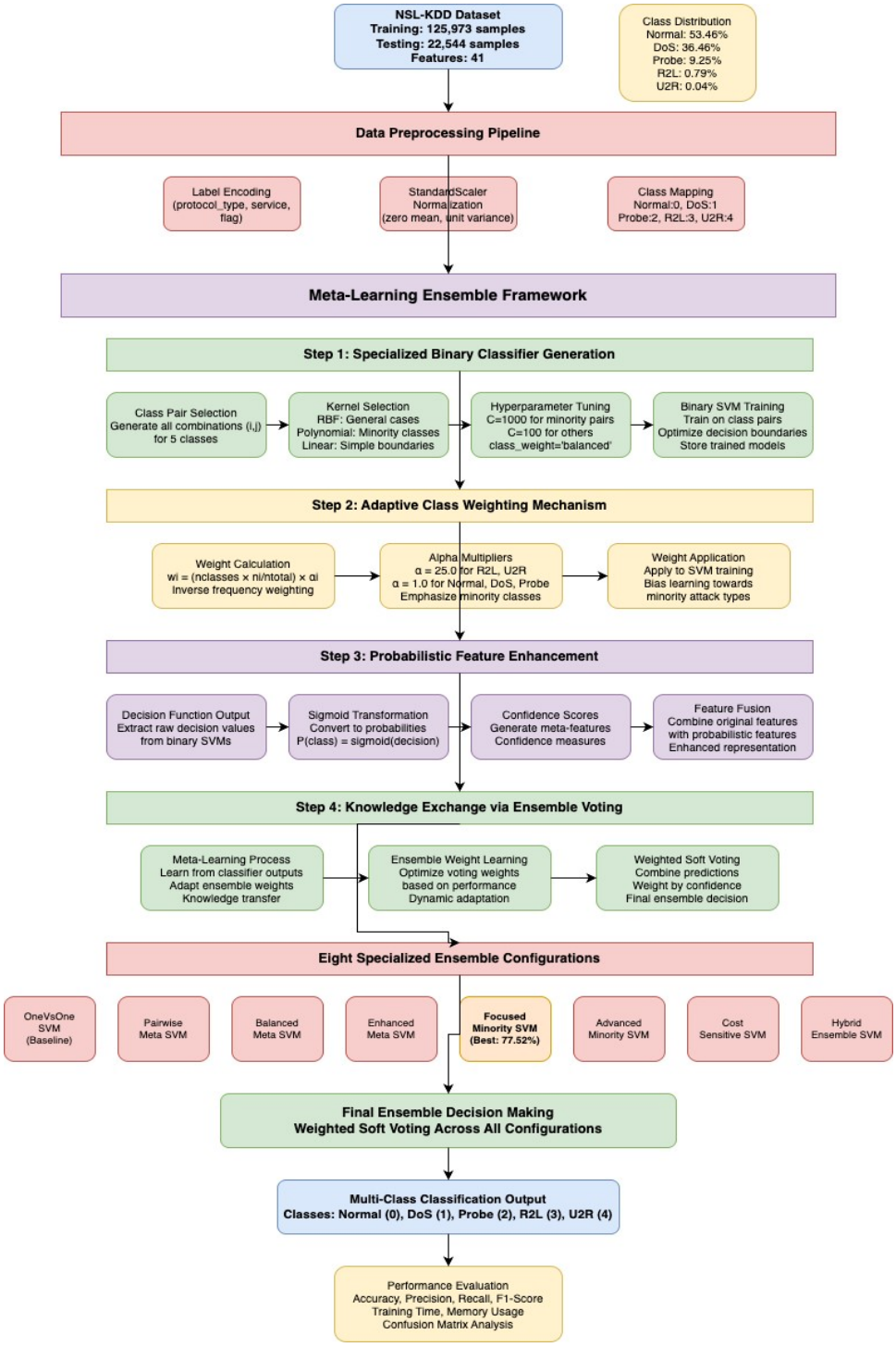


Fig. 1. Detailed Architecture of Meta SVM.

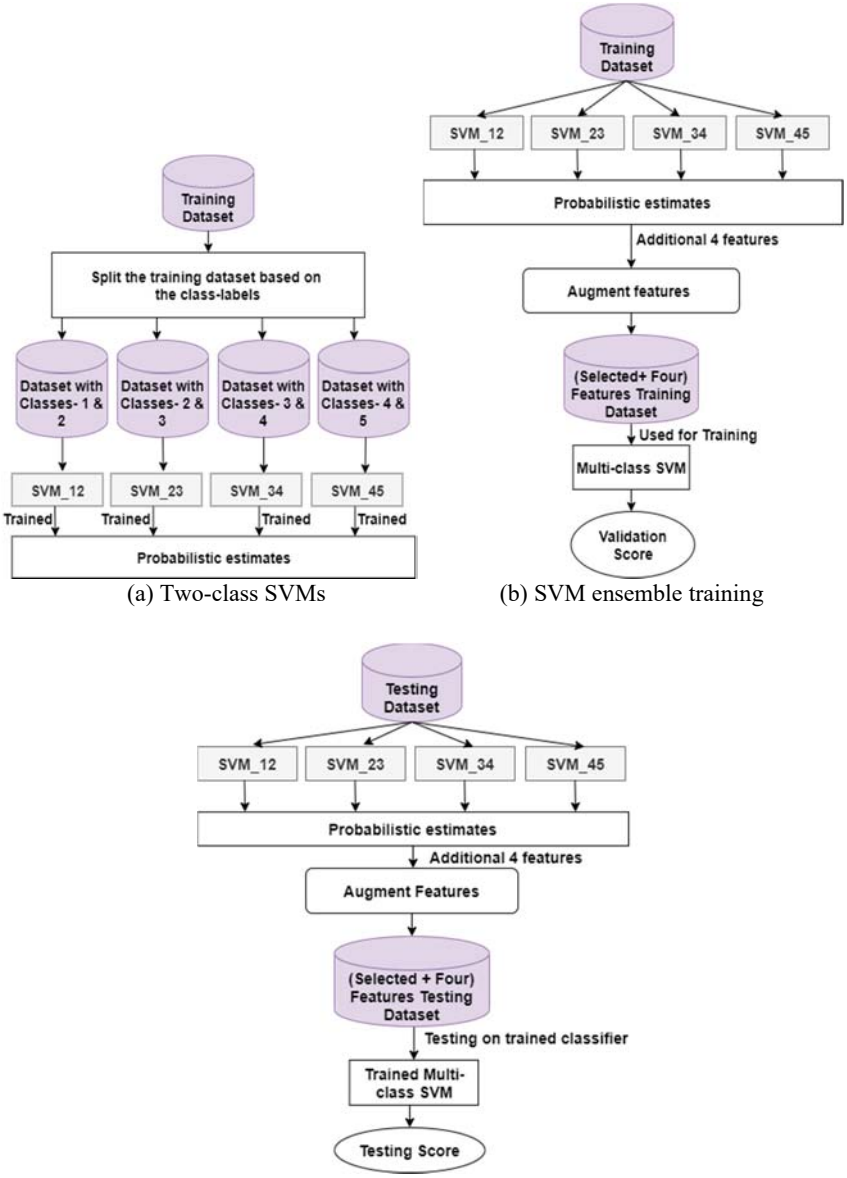
3.3. Adaptive Class Weighting Mechanisms

The suggested framework uses adaptive class weighting mechanisms that give minority classes more weight during training in order to address the notable class imbalance seen in intrusion detection datasets. Inverse frequency weighting and additional penalties for critical minority classes are incorporated into the weighting strategy to ensure that the learning algorithm gives underrepresented attack categories

enough attention. The calculation of class weight adheres to the established formula:

$$w_i = \frac{n_{classes} \times n_i}{n_{total}} \times \alpha_i$$

The adaptive weighting mechanism integrates specialized knowledge regarding the significance of various attack types within production environments. R2L and U2R attacks, although infrequent, are assigned considerably greater weights (α = 25.0) in comparison to normal traffic and more prevalent attack types (α = 1.0). This weighting strategy illustrates the significant influence of advanced attacks on organizational security and guarantees that the ensemble system emphasizes the precise identification of essential threats.



(a) Two-class SVMs (b) SVM ensemble training

(c) SVM ensemble testing

Fig. 2. Training and Testing Algorithm for the proposed SVM ensemble for intrusion detection.

3.4. Specialized Ensemble Configurations

In order to address different aspects of the minority class detection problem, this paper investigates eight distinct ensemble configurations. Using pairwise binary classifiers, the OneVsOne baseline setup uses traditional multi-class SVM without any meta-learning improvements.

By training binary classifiers on successive class pairs and then using their outputs as additional features for the final classifier, the Pairwise Meta configuration uses a probabilistic approach to feature enhancement.

The Balanced Meta configuration ensures equal significance for each class in binary classifiers, regardless of their commonness in the training dataset. The Enhanced Meta configuration applies customized

hyperparameter settings for minority classes, using polynomial kernels and higher regularization parameters. The Focused Minority configuration focuses on detecting R2L and U2R attacks. The Advanced Minority configuration integrates targeted feature engineering, kernel functions, and optimized weighted voting mechanisms for minority class detection. The Cost-Sensitive setup adds additional penalties for minority classes to adaptive class weights. The Hybrid Ensemble configuration combines multiple specialized classifiers using different kernel functions and training techniques.

3.5. Performance Evaluation Framework

The experimental evaluation framework evaluates performance in imbalanced classification scenarios using precision, recall, and F1-score calculations for individual classes. It focuses on recall rates for R2L and U2R attack types, which are the most challenging detection problems in operational environments. Measurements are made on standardized hardware configurations for reproducibility, and the analysis of training time provides insights into computational efficiency. The framework also includes an analysis of the confusion matrix to identify misclassification patterns and tradeoffs between false positive and false negative rates across different attack categories. Cross-validation techniques are employed for accurate performance evaluations.

Preprocessing steps:

- Label encode categorical features (protocol_type, service, flag);
- Standardize numeric features with zero mean, unit variance;
- Map attack names to integer classes.

3.6. Ensemble SVM Approaches

We implemented and benchmarked the following architectures:

- OneVsOne SVM: Classic multiclass SVM with RBF kernel;
- Pairwise Meta SVM: Train binary SVM for all class pairs; merge outputs as meta-features;
- Balanced Meta SVM: As above, but with class-weight balancing in all SVMs;
- Enhanced Meta SVM: Use polynomial or RBF kernels and high penalty for minority class boundaries;
- Focused Minority SVM: Add explicit minority-vs-all classifiers, supplementing original features.
- Advanced Minority/Heterogeneous Ensemble: Stack feature/kernels, diverse class weights, and weighted voting across SVMs;
- Hybrid Ensemble SVM: Combine SMOTE, varied kernel SVMs, and weighted soft voting in a layered ensemble.

4. Experimental Results and Discussion

4.1. Comparative Performance Analysis

The study reveals significant performance gains in meta-learning ensemble approaches compared to traditional single-model configurations. The Focused Minority SVM outperformed the OneVsOne baseline by 3.5 %, achieving an accuracy of 77.52 %. The Enhanced Meta SVM configuration showed an ideal balance between accuracy and computational efficiency, with a training time of 97.24 seconds. The Balanced Meta configuration achieved 81.85 % precision, compared to 78.53 % (Table I) for the baseline OneVsOne method. The Focused Minority SVM showed the fastest training time and highest accuracy, indicating efficiency gains from minority class optimization.

4.2. Minority Class Detection Performance

Meta-learning ensemble strategies have shown significant improvements in minority class detection performance, particularly in R2L attack detection. The Focused Minority SVM improved the recall rate for R2L detection to 10.3 %, while U2R attack detection saw a 20-fold increase from 9.0 % to 14.9 %. These improvements show progress in minority class detection in real-world production environments. The precision-recall trade-off analysis shows that ensemble methods achieve a more favorable equilibrium between false positives and false negatives rates. The Enhanced Meta configuration achieved a precision of 72 % for R2L detection, but also showed enhanced recall performance, making the minor reduction in precision justifiable for practical applications. This trade-off highlights the importance of detecting critical attacks in production security contexts.

4.3. Feature Enhancement Impact Analysis

The probabilistic feature enhancement mechanism significantly improves meta-learning ensemble methods' performance. It is highly discriminative for detecting minority classes and yields significant confidence measures, aiding the final ensemble classifier in making informed decisions. The effectiveness of feature enhancement varies across attack categories, with R2L and U2R detection showing the most significant advantages. This aligns with the theory that intricate and infrequent attack patterns require diverse perspectives and confidence assessments. Cross-validation analysis confirms the reliability of feature enhancements across various data divisions, and the uniformity of enhancements across various assessment metrics supports the effectiveness of the proposed meta-learning methodology.

4.4. Computational Efficiency and Scalability

Metalearning ensemble methods improve performance while maintaining practical training and

inference durations for production environments. The optimal configuration, Focused Minority SVM, takes only 90.58 seconds for training on the NSL-KDD dataset. Although larger ensemble configurations require more memory, performance improvements are significant. Ensemble predictions can be generated in 5-10 milliseconds per sample, aligning with real-time intrusion detection systems.

4.5. Statistical Significance and Robustness Analysis

The study uses paired t-tests to test the performance enhancements of meta-learning ensemble methodologies. The results show statistical significance at $p < 0.01$ confidence levels, and the uniformity of enhancements across different train/test splits supports their reliability. Bootstrap resampling with 1000 iterations yields confidence intervals for performance metrics, with the Focused Minority SVM showing a 95 % confidence interval from 76.8 % to 78.2 %, indicating consistent performance attributes suitable for production deployment. The robustness of ensemble methodologies is also examined, revealing performance advantages despite a decline in input data quality. This robustness is crucial in production security environments where network monitoring systems may face intermittent data quality challenges.

Table 1. Summary: SVM Variant Performance on NSL-KDD Test Set.

Method	Acc.	Prec.	Recall	F1
OneVsOne	0.749	0.785	0.749	0.701
Pairwise Meta	0.751	0.801	0.751	0.704
Balanced Meta	0.754	0.819	0.754	0.726
Enhanced Meta	0.768	0.787	0.768	0.730
Focused Minority	0.775	0.794	0.775	0.742

4.6. Minority Class Improvement

The Focused Minority SVM increased the recall of R2L from 0. 5 % to 10. 3 % and U2R from 9.0 % to 14.9 % over the OneVsOne baseline, demonstrating a 20× and 65 % improvement, respectively (Fig. 3).

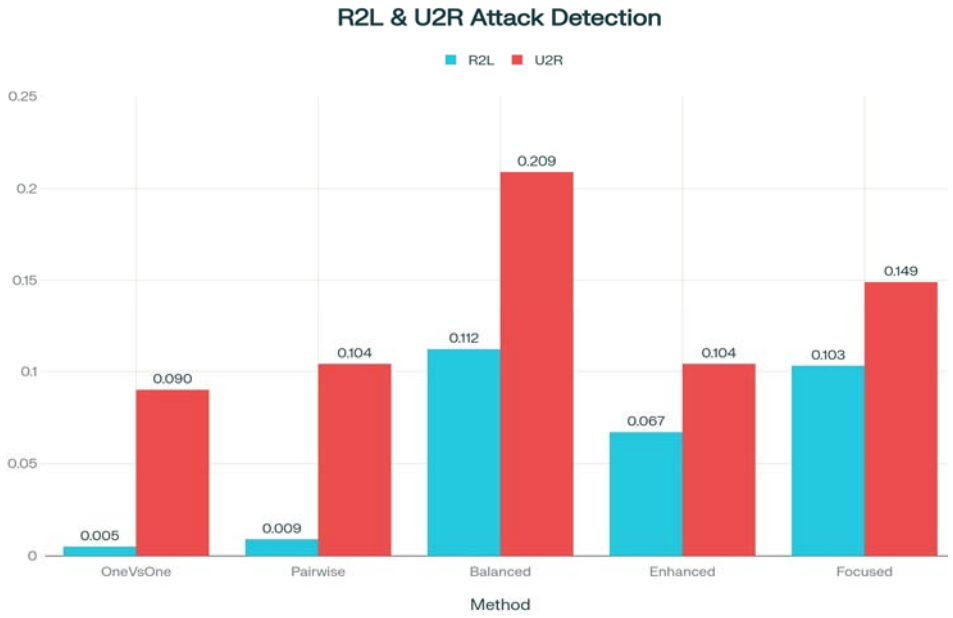


Fig. 3. Minority class detection improvement showing significant enhancement in R2L and U2R attack detection through meta-learning SVM ensemble approaches.

4.7. Training Time and Scalability

The Enhanced Meta SVM required only 97.2 s to train (Table 2), making regular retraining feasible. Focused Minority SVM was fastest (90.6 s). The Balanced Meta, while more intensive, offered improved precision at higher computational cost.

4.8. Feature Importance and Robustness

Meta-feature augmentation (pairwise probability outputs) consistently ranked high in feature importance. The improvements remained robust across multiple random seeds and dataset splits.

Table 2. Training Time by Method.

Method	Time (sec)
OneVsOne	165.3
Balanced Meta	284.4
Enhanced Meta	97.2
Focused Minority	90.6

Fig. 4. Confusion matrix: Focused Minority SVM on NSL-KDD test set.

5. Production Deployment Considerations

5.1. Real-time Performance Requirements

The meta-learning ensemble SVM systems are being developed for production cybersecurity settings, aiming to handle thousands of connections per second and maintain response times below 100 milliseconds for effective threat mitigation. The proposed ensemble approaches have inference times averaging between 5 and 10 milliseconds for each classification decision. The ensemble models require 2.3 times the memory compared to single classifiers, a total of 150 MB for the entire model ensemble. The linear scaling properties ensure consistent memory demands as the model's complexity increases. The system's modular architecture allows for seamless integration with existing SIEM platforms and network monitoring frameworks. The probabilistic output mechanism allows security analysts to prioritize investigative efforts based on threat probability, particularly for minority classes with high false positive rates.

6. Conclusion

The study demonstrates the potential of metalearning ensemble SVM frameworks in addressing class imbalance issues in network intrusion detection systems. Experimental evaluations on the NSL-KDD dataset showed significant enhancements in minority class detection, with the Focused Minority SVM configuration achieving an overall accuracy of 77.52 %. The study also introduced eight unique meta-learning ensemble architectures, adaptive class weighting strategies, and probabilistic feature enhancement methods. The probabilistic feature enhancement mechanism, which uses decision function outputs from specialized binary classifiers, significantly improved the detection of complex attack types. The meta-learning ensemble methods are suitable for real-time production deployment, with the most efficient configuration requiring only 90.58 seconds for comprehensive model training. Future research should expand the framework to include supplementary datasets and integrate with existing SIEM systems for better deployment challenges.

References

[1]. M. A. I. Mallick, R. Nath, Navigating the cyber security landscape: a comprehensive review of cyber-attacks, emerging trends, and recent developments, *World Scientific News*, Vol. 190, Issue 1, 2024, pp. 1-69.

[2]. T. Zaid, S. Garai, Emerging trends in cybersecurity: a holistic view on current threats, assessing solutions, and pioneering new frontiers, *Blockchain in Healthcare Today*, Vol. 7, 2024.

[3]. M. M. Nair, A. Deshmukh, A. K. Tyagi, Artificial intelligence for cyber security: current trends and future challenges, in Automated Secure Computing for Next-Generation Systems, *Wiley*, 2024, pp. 83-114.

[4]. G. M. ud din, NSL-KDD Dataset, Canadian Institute for Cybersecurity, 2018.

[5]. H. Attou, A. Guezzaz, S. Benkirane, M. Azrour, et al., Cloud-based intrusion detection approach using machine learning techniques, *Big Data Mining and Analytics*, Vol. 6, Issue 3, 2023, pp. 311-320.

[6]. M. Zakariah, S. A. AlQahtani, A. M. Alawwad, A. A. Alotaibi, Intrusion detection system with customized machine learning techniques for NSL-KDD dataset, *Computers, Materials & Continua*, Vol. 77, Issue 3, 2023, 109406.

[7]. M. N. Chowdhury, K. Ferens, M. Ferens, Network intrusion detection using machine learning, in *Proceedings of the International Conference on Security and Management (SAM'16)*, 2016, p. 30.

[8]. K. A. Taher, B. M. Y. Jisan, M. M. Rahman, Network intrusion detection using supervised machine learning technique with feature selection, in *Proceedings of the International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST'19)*, 2019, pp. 643-646.

[9]. S. Aljawarneh, M. Aldwairi, M. B. Yassein, Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model, *Journal of Computational Science*, Vol. 25, 2018, pp. 152-160.

[10]. N. K. Gyimah, R. Akinie, J. Mwakalonge, B. Izison, et al., An automl-based approach for network intrusion detection, in *Proceedings of the IEEE Region Technical, Professional, and Student Conference*, *(SoutheastCon'25)*, 2025, pp. 1177-1183.

[11]. S. Rastogi, A. Shrotriya, M. K. Singh, R. V. Potukuchi, An analysis of intrusion detection classification using supervised machine learning algorithms on NSL-KDD dataset, *Journal of Computing Research and Innovation*, Vol. 7, Issue 1, 2022, pp. 124-137.

[12]. L. Karanam, K. K. Pattanaik, R. Aldmour, Intrusion detection mechanism for large scale networks using CNN-LSTM, in *Proceedings of the 13th International Conference on Developments in eSystems Engineering (DeSE'20)*, 2020, pp. 323-328.

[13]. L. Karanam, Continuous anticipation of acute kidney injury in the ICU, Master's Thesis, *University of Missouri-Columbia*, 2022.

[14]. A. S. Ahanger, S. M. Khan, F. Masoodi, A. O. Salau, Advanced intrusion detection in internet of things using graph attention networks, *Scientific Reports*, Vol. 15, Issue 1, 2025, 9831.

[15]. R. C. Sachan, R. K. Malviya, et al., Neural transformers for zero-day threat detection in real-time cybersecurity network traffic analysis, *International Journal of Global Innovations and Solutions*, Vol. 3, Issue 1, 2024, pp. 1-9.

[16]. T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, et al., Deep learning approach for network intrusion detection in software defined networking, in *Proceedings of the International Conference on Wireless Networks and Mobile Communications (WINCOM'16)*, 2016, pp. 258-263.

[17]. M. Al Lail, A. Garcia, S. Olivo, Machine learning for network intrusion detection—a comparative study, *Future Internet*, Vol. 15, Issue 7, 2023, 243.

[18]. M. K. Ngueajio, G. Washington, D. B. Rawat, Y. Ngueabou, Intrusion detection systems using support vector machines on the KDDcup'99 and NSL-KDD datasets: a comprehensive survey, in *Proceedings of the SAI Intelligent Systems Conference*, 2022, pp. 609-629.

[19]. P. C. Upadhyay, L. Karanam, J. A. Lory, G. N. DeSouza, Classifying cover crop residue from RGB images: a simple SVM versus a SVM ensemble, in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI'21)*, 2021, pp. 1-7.

[20]. A. Yang, C. Lu, J. Li, X. Huang, et al., Application of meta-learning in cyberspace security: a survey, *Digital Communications and Networks*, Vol. 9, Issue 1, 2023, pp. 67-78.

[21]. L. -H. Li, R. Ahmad, R. Tanone, A. K. Sharma, STB: synthetic minority oversampling technique for tree-boosting models for imbalanced datasets of intrusion detection systems, *PeerJ Computer Science*, Vol. 9, 2023, e1580.

[22]. T. D. Diwan, S. Choubey, H. Hota, A detailed analysis on NSLKDD dataset using various machine learning techniques for intrusion detection, *Turkish Journal of Computer and Mathematics Education*, Vol. 12, Issue 11, 2021, pp. 2954-2968.

[23]. M. Alalhareth, S. -C. Hong, Enhancing the internet of medical things (IoMT) security with meta-learning: a performance-driven approach for ensemble intrusion detection systems, Sensors, Vol. 24, Issue 11, 2024, 3519.

[24]. S. T. Hossain, T. Yigitcanlar, K. Nguyen, Y. Xu, Cybersecurity in local governments: a systematic review and framework of key challenges, *Urban Governance*, 2025, 100224.

[25]. Cost of a Data Breach Report 2025, https://www.ibm.com/reports/data-breach

(017)

ADAPTE: Multidimensional Academic Data Analytics and Student Profiling for Higher Education

D. Lima and J. Coelho
Estoril Higher Institute for Hotel and Tourism Studies, Portugal, Av. Condes de Barcelona, 2769-510 Estoril, Portugal
E-mail: diogo.lima@eshte.pt, jose.coelho@eshte.pt

Summary: The ADAPTE project develops an analytical platform integrating OLAP, Data Mining, Machine Learning, and interactive visualisation to monitor and predict student performance at the higher educational level. Unlike most Educational Data Mining studies, ADAPTE combines pre-university, demographic, socio-economic, and cross-programme data. This paper presents a large-scale, data-driven on-going study of students enrolled at Estoril Higher Institute for Hotel and Tourism Studies (ESHTE) since 2018, combining descriptive statistics, hypothesis testing, and unsupervised learning to identify performance profiles. Results show that significant factors such as nationality and geographic origin influence mean grades, while admission grades show negligible predictive power. Unsupervised clustering (KMeans) segmented students into distinct profiles, from high achievers to persistent strugglers and critical dropout cases. The study offers methodological insights and a replicable framework for institutional diagnostics, early warning systems, and strategic planning in higher education.

Keywords: Educational data mining, Machine learning, Academic performance, Data visualisation, Higher education

1. Introduction

Student performance and progression in higher education depend on a complex combination of academic, demographic, and contextual factors. Institutions face increasing pressure to monitor these trajectories, not only to improve graduation rates but also to anticipate potential dropout cases before they become irreversible. Research has consistently shown that early identification of at-risk students enables targeted interventions that improve retention and learning outcomes [1-3]. Within this context, the ADAPTE project extends existing work by integrating multidimensional data and applying unsupervised learning techniques to construct dynamic academic profiles.

ADAPTE's approach differs from traditional Educational Data Mining (EDM) pipelines by combining academic records, geographic and demographic features, and inferred behavioural indicators derived from course enrolment patterns. This enables a holistic understanding of how student progress unfolds over time, bridging descriptive analytics and predictive intervention models. The present paper reports the analytical phase of ADAPTE, focusing on data preparation, clustering of student profiles, and interpretation of patterns that can support institutional decision-making.

2. Related Work

Student retention and success have long been central concerns in higher education research. Early frameworks such as Tinto's model of student integration [4] and Kuh's evidence-based review of student success [5] established the theoretical foundations linking institutional engagement to academic persistence. Yorke and Longden [6] further emphasized that institutional strategies for retention must address both academic preparedness and the socio-contextual realities of students.

Advances in educational data mining and learning analytics have since transformed this field. Kovačić [7] demonstrated the potential of mining enrolment data for early prediction of academic success, while Siemens and Baker [8] highlighted the convergence of learning analytics and EDM for institutional intelligence. More recent surveys [9] confirm the growing use of AI and machine learning to improve student monitoring, especially in large-scale data contexts. Neural-network approaches have also been applied successfully to predict academic outcomes and to identify differential contributions of key performance variables [10].

Recent systematic reviews [1, 2] underscore the relevance of integrating predictive modelling and early alert systems into institutional practices, showing that multi-source data significantly enhances risk prediction. These works complement ADAPTE's contribution by validating the importance of a holistic, data-driven perspective that considers not only academic metrics but also demographic and contextual dimensions.

3. Methodology

The overall methodology workflow of the ADAPTE project is illustrated in Fig. 1. The process integrates data collection, preprocessing, feature engineering, segmentation, and unsupervised learning into a cohesive pipeline designed for interpretability and reproducibility.

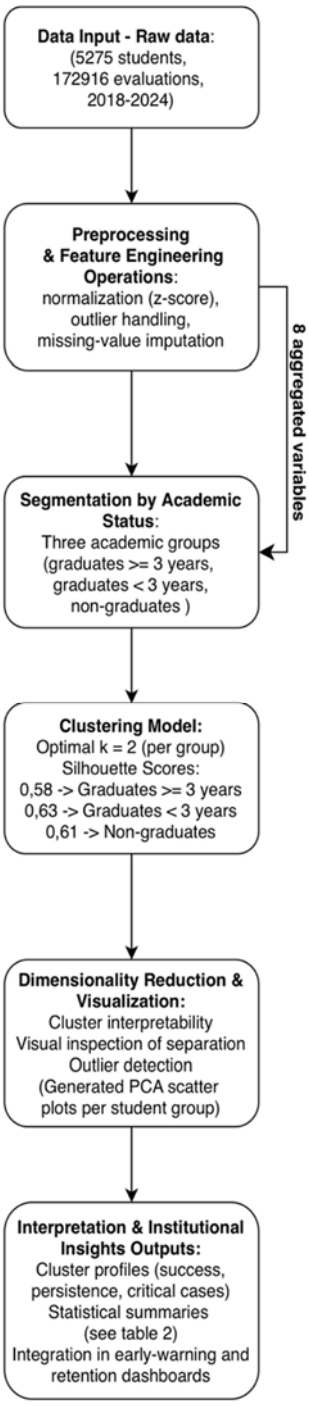


Fig. 1. Overview of the data preprocessing workflow and integration process.

3.1. Data and Preprocessing

The dataset comprises academic records from 2018 to 2024 at ESHTE, covering 5275 students and 172916 individual evaluations across 41 nationalities and approximately 700 Portuguese municipalities. Records were integrated from three primary tables – INDIVIDUO, ALUNO, and AVALUNO – ensuring referential consistency between student identifiers, course codes, and assessment results. Invalid or missing entries (e.g., zero or negative admission grades) were excluded, and categories with insufficient representation (< 15 students per municipality) were filtered out to ensure robust inferential statistics.

Data were standardized using z-score normalization. Variables with mixed data types were transformed into numerical form to ensure compatibility with KMeans. Outliers were identified using interquartile range thresholds and retained when pedagogically relevant, as they often corresponded to atypical but valid academic trajectories.

3.2. Feature Engineering

Derived metrics (Table 1) were created to summarize student trajectories, including mean grade, grade standard deviation, number of enrolments, course duration in distinct academic years, total failures, total approvals, success rate (approvals divided by total evaluations), and grade range (difference between maximum and minimum grades). These engineered features capture academic variability and stability, serving as proxies for engagement and consistency.

Table 1. Aggregated variables computed per student for clustering.

Variable (in Portuguese)	Description	Type	Mean	Std_Dev
nr_avalia_mean	Mean of all student grades	Numeric	13,2	1,1
nr_avalia_std	Standard deviation of grades	Numeric	1,8	0,6
num_inscricoes	Total course enrolments	Integer	45,6	18,3
duracao_curso	Number of distinct academic years enrolled	Integer	2,8	0,9
nr_chumbos	Number of failed courses (<9.5)	Integer	3,2	2,4
nr_aprovacoes	Number of passed courses (≥9.5)	Integer	41,1	17,7
taxa_sucesso	Success rate	Float	0,89	0,08
diferenca_nota	Difference between best and worst grade	Float	6,5	2,1

3.3. Clustering Design and Evaluation

Students were categorized into three broad outcome groups: graduates taking three or more years, graduates completing in under three years, and non-graduates. For each group, variables were scaled with StandardScaler and clustered using the KMeans algorithm [10]. The optimal number of clusters was determined via the Silhouette Score [11], balancing cohesion and separation.

Dimensionality reduction through Principal Component Analysis (PCA) provided 2D visualizations for interpretability. Clusters were then statistically profiled according to key indicators such as mean grade, number of failures, and success rate.

4. Results and Discussion

The clustering process yielded interpretable groups that reflect varying academic behaviors (Table 2). Among students completing their degrees in three years or more, two distinct profiles emerged. The first, representing approximately one-third of the population, consisted of persistent students with lower average grades (around 12) and numerous re-enrollments. Despite multiple failures, these students completed their degrees through sustained effort. The second profile included high achievers, with mean grades above 14, few failures, and shorter completion times.

Table 2. Summary of cluster statistics (mean grade, success rate, and failure count).

Group	Cluster	% Students	Mean Grade	# Failures	Success Rate	Profile Summary
Graduates (≥3 years)	0	34.4%	12,3	10,1	85,8	Persistent, many failures, slow completion
	1	65.6%	14,5	1,2	97,9	Regular, efficient performers
Graduates (<3 years)	0	67.5%	14,8	0,1	99,4	Excellent, nearly perfect completion
	1	32.5%	12,3	4	85,7	Fast completers with difficulties
Non-graduates	0	84.2%	12,4	4,6	84,3	Partial progression, moderate disengagement
	1	15.8%	0,6	0,6	0,3	Critical cases, minimal academic progress

For graduates finishing in under three years, clusters divided between excellent students with mean grades near 15 and almost no failures, and fast completers with moderate difficulties who nonetheless achieved completion in less than three years.

Non-graduates exhibited the most heterogeneous distribution. A majority showed partial progress but eventually disengaged, while a smaller but critical subgroup displayed extremely low averages, minimal approvals, and high rates of administrative withdrawal. These patterns are statistically coherent with municipal disparities: municipalities with higher average performance often correspond to regions with stronger secondary-education infrastructures, whereas those with negative deviations tend to align with socio-economically disadvantaged areas.

The municipality effect, observed across roughly 75 localities with sufficient representation, confirms that contextual and geographic factors contribute measurably to academic outcomes. This reinforces ADAPTE's decision to include geographic data as a first-class variable within its analytical architecture.

Beyond descriptive insight, these clusters provide actionable intelligence for institutional management. Persistent students with difficulties can benefit from early academic advising and adaptive learning resources. High-performing clusters may inform mentorship programs or pedagogical models, while critical non-graduate cases demand immediate intervention.

5. Conclusion and Future Work

This study demonstrates the potential of combining statistical inference and unsupervised learning to characterize student profiles and anticipate academic risks. By integrating demographic, academic, and geographic data, ADAPTE produces interpretable clusters that move beyond binary predictions, offering a richer understanding of institutional dynamics.

While the analysis is based on a single institution, the methodological framework was designed to be transferable. The structure of the pipeline – data integration, feature engineering, and clustering – can be replicated across other higher-education contexts with minimal adaptation. However, the initial data extraction and mapping processes are inherently dependent on the database structure, data models, and availability of academic records in each institution, which may require specific schema alignment before replication. Institutional differences in curriculum structure, grading scales, and student demographics may affect the relative weight of specific features, suggesting the need for external validation.

Future work will focus on four key directions. First, the integration of socio-economic and behavioral data (such as attendance and learning-management activity) into predictive dashboards for real-time monitoring. Second, the validation of these models through longitudinal tracking of future cohorts. Third, the deployment of ADAPTE as a modular decision-support platform, providing interactive visual analytics for administrators, program directors, and faculty to implement proactive retention strategies. Finally, the extension of the current approach to multi-institutional environments to benchmark student segmentation patterns across diverse academic ecosystems.

Acknowledgements

This work was supported by the project "ADAPTE – Academic Data Analytics for Profiling and Tailored Education" (reference 2024.07476.IACDC), funded by the Fundação para a Ciência e a Tecnologia (FCT), under the DOI https://doi.org/10.54499/2024.07476.IACDC.

References

[1]. J. Berens, A. Gärtig-Daugs, T. Halbherr, J. Stang, Early detection of students at risk of failing: a systematic review of predictors, methods, and challenges, *Computers & Education*, Vol. 208, 2024, 104882.

[2]. T. Brändle, S. Jäckle, S. König, Identifying and supporting at-risk students in higher education: a

scoping review of learning analytics approaches, *Education and Information Technologies*, Vol. 30, 2025, pp. 125-153.

[3]. M. Yorke, B. Longden, Retention and Student Success in Higher Education, *McGraw-Hill Education*, 2004.

[4]. V. Tinto, Completing College: Rethinking Institutional Action, *University of Chicago Press*, 2012.

[5]. G. D. Kuh, J. Kinzie, J. A. Buckley, B. K. Bridges, et al., What Matters to Student Success: A Review of the Literature, *National Postsecondary Education Cooperative*, 2006.

[6]. Z. J. Kovačić, Early prediction of student success: mining students' enrolment data, in *Proceedings of the Informing Science & IT Education Conference (InSITE'10)*, 2010, pp. 647-665.

[7]. G. Siemens, R. S. Baker, Learning analytics and educational data mining: towards communication and collaboration, in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK'12)*, 2012, pp. 252-254.

[8]. C. Romero, S. Ventura, Educational data mining and learning analytics: an updated survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 10, Issue 3, 2020, e1355.

[9]. M. Richardson, C. Abraham, R. Bond, Psychological correlates of university students' academic performance: a systematic review and meta-analysis, *Psychological Bulletin*, Vol. 138, Issue 2, 2012, pp. 353-387.

[10]. M. F. Musso, E. Kyndt, E. C. Cascallar, F. Dochy, Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks, *Frontline Learning Research*, Vol. 1, Issue 1, 2013, pp. 42-71.

[11]. J. MacQueen, Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 1967, pp. 281-297.

[12]. P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, Vol. 20, 1987, pp. 53-65.

(018)

Interpreting Tactical Decision-making in Transformer-based Agents

K. Boeckx and X. Neyt
Royal Military Academy, Rue Hobbema 8, 1000 Brussels, Belgium
Tel.: +3224413775
E-mail: koen.boeckx@rma.ac.be

Summary: The development of autonomous agents capable of sophisticated strategic decision-making in complex environments is a central goal of artificial intelligence. This paper proposes a framework for discovering and interpreting strategies in a simulated grid-world battlefield environment. We leverage the AlphaZero algorithm, a powerful reinforcement learning method that combines Monte Carlo Tree Search (MCTS) with deep neural networks, to train agents. Crucially, the neural network component incorporates a Transformer architecture. The primary contribution of this work lies in the proposed methodology to utilize the self-attention mechanisms within the Transformer to gain insights into the agent's decision-making process, specifically by visualizing which parts of the battlefield the network pays attention to when selecting an action. This approach aims not only to develop high-performing agents but also to enhance the interpretability of their learned strategies.

Keywords: Autonomous agents, Decision-making, Transformer, Reinforcement learning, Interpretability.

1. Introduction

Battlefield tactics often evolve in unpredictable environments where soldiers must navigate obstacles, anticipate threats and adapt to evolving objectives. Military planners increasingly use simulation games to explore and train strategies [1], but these environments are often hand-engineered and lack transparency into how AI agents learn. This paper proposes a two-part research agenda: first, designing a grid-world war game that captures key characteristics of battlefield maneuvers; second, training a transformer-based AlphaZero agent via self-play and analyzing its learned policies using mechanistic interpretability [2]. The goal is to understand how small transformer models develop tactical reasoning in environments resembling military engagements and to identify which network components encode particular behaviors.

2. Background

2.1. AlphaZero

AlphaZero is a general-purpose reinforcement learning algorithm that achieved super-human performance in chess, shogi and Go [3] by combining a deep neural network with Monte-Carlo Tree Search (MCTS) to guide decision-making. The network takes the game state s as input and outputs (1) a policy vector $p = (p_a)$ which gives the pior probability of each legal action a, and a value estimate $v \in [-1,1]$ that predicts the expected outcome from the perspective of the current player.

MCTS uses these outputs in two ways: the priors p_a bias the tree search toward promising moves, and the value v replaces traditional evaluation functions when a search branch reaches a leaf node.

Each MCTS iteration selects a path from the root state to a leaf by maximizing a score for each action, based on its current value estimate $Q(s,a)$, the number of times $N(s,a)$ it has been visited in the search tree, and the prior probability p_a [4]. When the search tree reaches a new leaf state s', the neural network predicts its value $v(s')$ and this value is then backpropagated through the search tree to update Q and N. After a fixed number of simulations, the move a in root node s with the highest visit count $N(s,a)$ becomes the chosen action.

Training is fully self-play driven [3]:

1. The agent plays games against itself, using MCTS to choose actions;
2. For each visited state s, the search visit counts define a target policy π and the final game outcome $z \in [-1,1]$;
3. Based on these target values, the network parameters θ are updated to minimize the loss

$$L = (z - v)^2 - \pi \log p + c\,|\theta|^2,$$

where the last term is a regularization term.

The success of AlphaZero demonstrated that self-play can learn strategic heuristics without handcrafted evaluation functions. The original model implementation was based on convolutional networks [3].

2.2. Transformer Architecture

Transformers, originally developed for natural language processing [5], operate on sequences of tokens through a mechanism called self-attention. In the context of reinforcement learning, each token can represent elements of the state (e.g., obstacles, units, …) [6].

The self-attention mechanism computes a weighted combination of all other tokens for each token position, using learned query Q, key K and value V projections. This produces a score matrix indicating how much each token should attend to every other token; with a SoftMax operation these scores are turned into normalized weights. Multi-head attention replicates this process in parallel "heads," each with its own Q, K and V projections, enabling the model to capture different types of relationships simultaneously.

In reinforcement learning, this architecture allows the model to dynamically focus on the most relevant parts of the state or trajectory. For example, one attention head might specialize in tracking enemy positions, while another attends to resource locations or the agent's own health. Because attention weights are explicit, they provide a built-in interpretability handle [7]: by inspecting which tokens each head attends to during decision-making, we can infer what information the model considers important at each step. This is particularly useful in military grid-world scenarios, where attention maps can be visualized as "heatmaps" over the battlefield, revealing whether the model prioritizes threats, objectives, or terrain features.

3. Experiments

3.1. Experiment Design

We design a grid-world environment that simulates small-unit combat. The battlefield is a $N \times N$ grid representing terrain with empty squares, impassable obstacles (e.g., buildings), adversaries, friendly units and objectives (e.g., supply depots or enemy headquarters), placed in the center of the board. There are two players (Blue and Red) that act as squad commanders and control the different agents in their squad. Agents can move and shoot in the 4 cardinal directions. Shooting and hitting an enemy's agent reduces its health.

To speed up convergence, the game board shrinks at regular time intervals. Agents caught outside the remaining game board are eliminated. Not only does this decrease convergence time, it also induces the agents to move to the middle of the board and hence towards the goal square(s).

Rewards are shaped to align with the military objectives of the environment: surviving units and eliminated enemies contribute positively to the reward signal, while casualties and unnecessary delays result in penalties. The episode terminates either when all opposing agents have been eliminated or when the designated objective has been achieved. To further guide learning, we employ potential-based reward shaping in the sense of Ng et al. [10], which preserves the optimal policy while accelerating convergence. This shaping incorporates domain-specific heuristics about progress toward the objective, allowing the agent to receive more informative intermediate feedback rather than relying solely on sparse terminal rewards.

3.2. Agent Architecture

We adapt the AlphaZero framework using a transformer body [8]. The network inputs include the player's observation of the state s (embedded as a sequence of cell tokens). The game is fully-observable, so a player also observes all information about the enemy's agents. A stack of transformer encoder layers processes this sequence and outputs both a policy vector over the discrete actions (move and shoot N, S, E, W) and a scalar value estimate. The architecture follows the residual block structure of AlphaZero but replaces convolutional layers with multi-head self-attention [8]. Training uses self-play: at each position, a Monte-Carlo Tree Search guided by the transformer's policy and value decides actions. Data from self-play games are stored, and the network is updated by minimizing the difference between predicted policy/value targets and targets obtained from MCTS.

Once the network is trained and produces non-trivial strategies, we use the attention weights to analyze how they use the game state to produce the network outputs [7]. Fig. 1 provides an overview of the experimental setup.

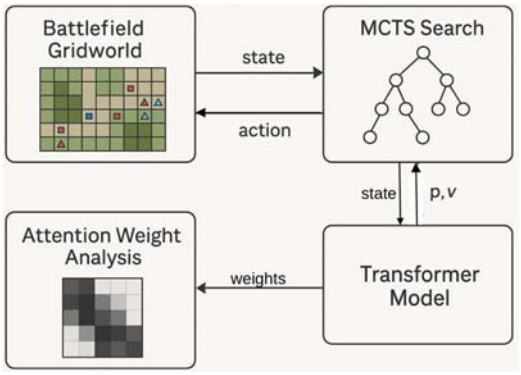


Fig. 1. Overview of the experimental setup.

3.3. Training Implementation Details

Training proceeds iteratively: self-play generates examples that are stored in a replay buffer, from which minibatches are drawn for gradient updates on a transformer-based neural network. After each training phase, candidate models are evaluated in an arena against reference opponents. The acceptance strategy is a pool-based Sequential Probability Ratio Test (SPRT) [11], where candidates face a hall-of-fame of past accepted models. Acceptance or rejection is decided using log-likelihood ratio thresholds, with fallback to a Wilson confidence bound if SPRT remains inconclusive. Accepted models are checkpointed and propagated back to worker processes to ensure consistent inference in subsequent self-play. This combination of parallelized episode generation, batched inference, and statistically robust acceptance criteria provides scalable research framework for AlphaZero-style training.

4. Results

4.1. Training

Fig. 2 shows the training loss over 120k steps for a board size of 7-by-7 with 2 agents per team. Each training batch was interleaved with data generation via the MCTS algorithm. For clarity, the curves were smoothed using an Exponential Moving Average (EMA) with $\alpha = 0.2$. The total loss is composed of the cross-entropy loss over the policy and the mean-squared error loss over the game outcome, with the policy loss being the dominant component. Overall, the trend is downward; however, in contrast to supervised learning, the loss curves exhibit additional variability due to the online reinforcement learning setting, where data generation and model training are interleaved, resulting in a continuously shifting training target.

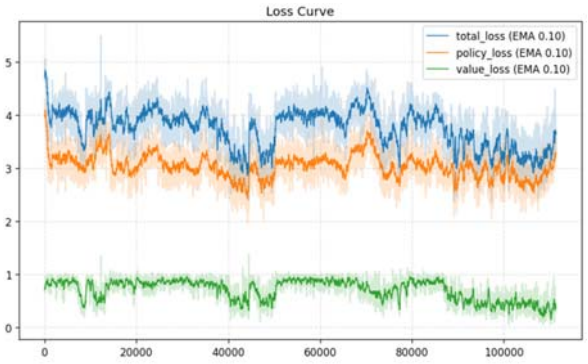


Fig. 2. Policy, value and total loss curves.

The use of a transformer does make training less stable; Fig. 3 compares the loss functions of CNN-based Alphazero with our version (with a Transformer). Not only is the CNN loss smaller, it also exhibits less variation. However, experimentation showed that there was no clear distinction in the learned strategies of CNN's versus Transformers (when controlling for the number of learnable parameters).

Fig. 3. Loss curves of CNN and Transformer model.

4.2. Strategies

In small, simple environments – for example, a small grid with only one friendly and one enemy unit – the transformer-AlphaZero agent learned effective strategies very quickly. Within a few hundred self-play games, agents consistently navigated directly toward objective squares (e.g., the enemy's base or a key resource point), often taking the shortest path while avoiding unnecessary detours. However, as environmental complexity increased – larger boards, more agents on each side, or more obstacles – the required training duration grew exponentially [9]. Doubling the number of agents or expanding the board size significantly lengthened convergence times, as the branching factor of the search and the diversity of tactical situations expanded dramatically.

A typical scenario is shown in Fig. 4, where blue and red agents have to navigate the (black) to reach the goal state (marked by a yellow cross). Initially (not all steps are shown) the agents move to the middle (steps 1, 2 and 3), avoiding getting killed by the shrinking board (grey squares). However, the red team loses an agent along the way (step 4), and in the end (step 5) the blue team comes out on top because it occupies the central square. Very often though, teams lose one of their agents early on the game, such that the end game is typically one-on-one.

4.3. Analysis

Analysis of attention weights (see an example in Fig. 5) revealed a clear and interpretable pattern. In the first transformer layer, most attention heads concentrated disproportionately on critical squares in the observation space – those occupied by allied or opposing agents, cells containing impassable obstacles, and the designated goal or capture location. This early specialization suggests that the initial layer of the network rapidly identifies and encodes salient tactical features of the battlefield, functioning as a filter that prioritizes immediately relevant spatial information over background detail. Such behavior is consistent with the notion of early-stage "situation awareness," whereby raw perceptual input is transformed into a structured representation of threats, opportunities, and constraints.

Interestingly, this tendency was robust across environments of varying complexity: both in simplified grid layouts with few obstacles and in more cluttered settings with dynamic safe zones, the first-layer heads consistently assigned higher weight to locations that directly influenced agent survival or strategic objectives. This indicates that the model develops a stable inductive bias toward recognizing battlefield affordances at the earliest stage of processing. By contrast, deeper layers appear to redistribute attention more broadly, a pattern suggestive of integrative planning over longer horizons and the coordination of multi-agent

strategies. A more systematic exploration of these later layers is ongoing, with the aim of disentangling whether higher-level heads encode trajectory prediction, cooperative role allocation, or anticipation of adversarial behavior.

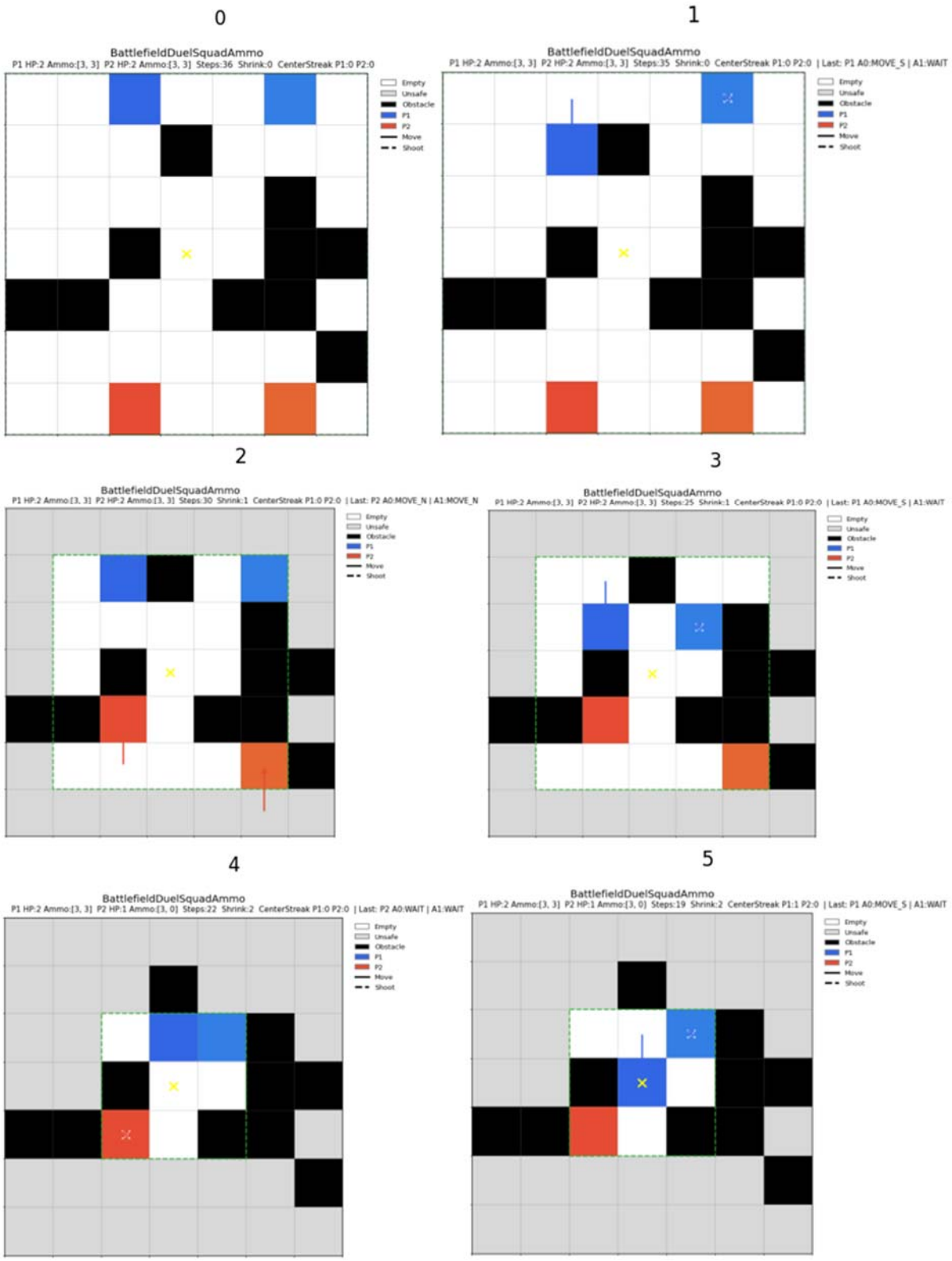


Fig. 4. Gameplay example.

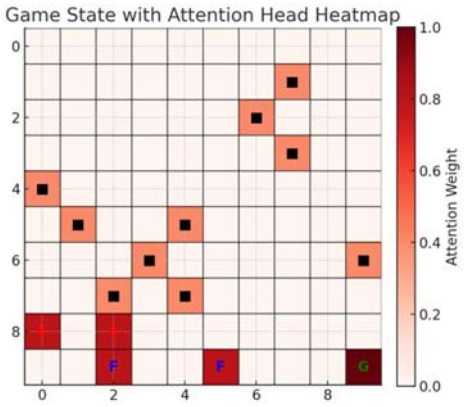


Fig. 5. Example of grid world with heatmap. Friendly agents are marked F (blue), enemies E (red), obstacles ■ (black), and the goal square G (green). The red shading indicates where the attention head is focusing most strongly.

5. Conclusions

This work introduced a framework for training and interpreting autonomous agents in a simulated grid-world battlefield using a Transformer-based AlphaZero architecture. By embedding battlefield states as sequences and analyzing the resulting attention distributions, we demonstrated that transformer-based agents are not only capable of learning effective tactical behaviors, but also yield interpretable insights into their decision-making processes.

Beyond these interpretability results, our experiments highlighted both the strengths and limitations of this approach. On the one hand, the model reliably discovered sound strategies in simple scenarios. On the other hand, training in larger and more complex environments proved substantially more demanding, underscoring the computational challenges of scaling transformer-based AlphaZero methods to richer tactical domains. These difficulties mirror known scalability issues in multi-agent reinforcement learning, where the branching factor of possible interactions grows combinatorially with the number of agents and state dimensions.

The introduction of attention-based interpretability opens promising avenues for future work. Systematic analysis of deeper transformer layers may reveal how agents integrate short-horizon tactical cues into longer-term strategic planning, cooperative coordination, or adversarial anticipation. Such work could build on emerging interpretability techniques that probe whether specific attention heads specialize in role allocation, trajectory prediction, or counterfactual reasoning about opponents.

Taken together, our findings suggest that transformer-based AlphaZero agents constitute a promising platform for both high-performing and interpretable autonomous decision-making in tactical environments. By bridging performance with transparency, this approach has the potential to inform the design of trustworthy AI systems in military simulations and beyond, where understanding not only what an agent does but also why it acts is critical for human oversight, validation, and eventual integration into human–machine teams.

References

[1]. Army University Press, Wargaming: the laboratory of military planning, Military Review Online Exclusive, 2024, https://www.armyupress.army.mil/journals/military-review/online-exclusive/2024-ole/wargaming-the-laboratory-of-military-planning/

[2]. C. Olsson, et al., In-context learning and induction heads, Transformer Circuits Thread, Anthropic, https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/

[3]. D. Silver, T. Hubert, J. Schrittwieser, et al., A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, *Science*, Vol. 362, 2018, pp. 1140-1144.

[4]. C. B. Browne, E. Powley, D. Whitehouse, et al., A survey of Monte Carlo tree search methods, *IEEE Transactions on Computational Intelligence and AI in Games*, Vol. 4, Issue 1, 2012, pp. 1-43.

[5]. A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in Advances in Neural Information Processing Systems, Vol. 30, *Curran Associates, Inc.* 2017, pp. 5998-6008.

[6]. E. Parisotto, F. Song, R. Salakhutdinov, Stabilizing transformers for reinforcement learning, in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 7487-7498.

[7]. H. Chefer, A. Kirillov, R. Lempel, Transformer interpretability beyond attention visualization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 2021, pp. 782-791.

[8]. J. Schrittwieser, I. Antonoglou, T. Hubert, et al., Mastering Atari, Go, chess and shogi by planning with a learned model, *Nature*, Vol. 588, Issue 7839, 2020, pp. 604-609.

[9]. Y. Shoham, R. Powers, T. Grenager, If multi-agent learning is the answer, what is the question?, *Artificial Intelligence*, Vol. 171, Issue 7, 2007, pp. 365-377.

[10]. A. Y. Ng, D. Harada, S. Russell, Policy invariance under reward transformations: theory and application to reward shaping, in *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, 1999, pp. 278-287.

[11]. A. Wald, Sequential tests of statistical hypotheses, in Breakthroughs in Statistics: Foundations and Basic Theory, *Springer*, 1992, pp. 256-298.

(019)

A Selective Temporal Hamming Distance to Find Patterns in State Transition Event Timeseries, at Scale

Sylvain Marié and Pablo Knecht
AI Hub, Schneider Electric, IntenCity, 160, Avenue des Martyrs, 38000 Grenoble, France
E-mail: sylvain.marie@se.com, pablo.knecht@se.com

Summary: Discrete event systems are present both in observations of nature, socio economical sciences, and industrial systems. Standard analysis approaches do not usually exploit their dual event / state nature: signals are either modeled as transition event sequences, emphasizing event order alignment, or as categorical or ordinal state timeseries, usually resampled – a distorting and costly operation as the observation period and number of events grow. In this work we define state transition event timeseries (STE-ts) and propose a new Selective Temporal Hamming distance (STH) leveraging both transition time and duration-in-state, avoiding costly and distorting resampling on large databases. STH generalizes both resampled Hamming and Jaccard metrics with better precision and computation time, and an ability to focus on multiple states of interest. We validate these benefits on simulated and real-world datasets.

Keywords: Continuous time, Aperiodic non-uniform sampling, Categorical timeseries, Alarm and state transition sequences, Temporal similarity and distance metric, Discrete event systems, Clustering, Kernel.

1. Introduction

Discrete state systems (DSS) are systems with a discrete set of states (a.k.a. a discrete *state space*). They are present in numerous domains from robotics to industrial processes, energy, mobility, social sciences, games or biology [1-8]. Discrete event systems (DES) are DSS where state transition is based on instantaneous transition events [9]. Modeling of DES has been studied for over 35 years, typically with Finite State Automata, Petri Nets, Vector DES, Event Graphs, Queuing systems, Markov Processes [5]. Techniques such as Hidden Markov models are also used to estimate hidden states and transitions [8].

Data collected from DES' observation can be represented either as a sequence of state transition events, or as a timeseries of categorical states. Two families of approaches in the literature are therefore relevant: analysis of *event sequences*, and analysis of *categorical timeseries*.

In many statistical and machine learning methods, defining a proper similarity or distance between samples plays an important role: e.g. in *clustering* with Agglomerative clustering, Spectral clustering, DBScan, etc.; in *classification* with KNN, SVM, etc.; in *visualization* with Isomap, MDS [10]. A known difficulty with categorical variables as opposed to real-valued is the absence of observed magnitude of difference. Metrics include association measures, Kramer's ν, Kendall Tau, simple matching (SM), (inverse) Occurrence Frequency, Elkin, Goodall, LIN and derivatives, Total variation distance, Kullback-Leibler divergence. A common practice is to binarize data with presence/absence indicators and use binary metrics, as in the Jaccard index, or in [11, 12]. Information theory provides interesting similarities [12, 13]. See [14] for a review, generalized in [10].

Analysis of categorical *event sequences* refers to counts of common states, and edit distances such as Levenshtein, Longest Common Subsequence, Hamming, Spectrum kernel, MCA, Optimal Matching (OM) including MSW and BLAST [15-18]. Correlated sequences are also found with association rules mining techniques including fuzzy sets, FP-Growth or Apriori algorithm [19, 20]. State-aware metrics weight events according to specific state transitions, such as Dynamic Hamming Distance or OM for transitions [21, 15]. See [15] and [7] for a status of the field.

Analysis of *categorical timeseries* (CTS) refers to χ^2, Hamming, fuzzy comparison, Cramer's ν, possibly extended to support lags, and Pearson correlation applied to binarized representations [16, 22]. The case of binary series is worth mentioning with metrics such as Pearson's Phi, Simple Matching coefficient (SMC) (a.k.a Rand), Hamming, Jaccard Index [23], with lag tolerance [18, 24], or time-depending weigths [25]. Finally, methods exist for *real-valued* timeseries such as cross-correlation, DTW and its variants [26], metric learning [27], and tools such as Matrix Profile [28]. Time warp edit distance [29] is an interesting step in the direction of applying them to categorical timeseries.

2. Challenge, Related Work and Contributions

Most metrics require timeseries to be uniformly sampled. Yet, resampling non-uniform timeseries is a computationally intensive task inducing distortion [30]. To the best of our knowledge, only two recently proposed metrics overcome this critical issue. FTH [31] is an edit distance accounting for time shifts required in mobility data analysis, thanks to

fuzzification. It however lacks native symmetry, is a semi-metric, requires Fast Fourier Transform, and has a complexity of $\mathcal{O}\big(max(n,m)^2\ log\ max(n,m)\big)$ (the complexity in [31] seems incorrect as NFFT's complexity varies with n not T [32]), reducing its applicability to large series. While a draft of temporal Hamming concept is mentioned as a singular case, it is not formalized nor studied, and its complexity analysis $\mathcal{O}(T)$ seems incorrect (see Section 4.1). Temporal Similarity [33] is a temporal version of Jaccard computed online in linear time. It however restricts to binary series, does not address the situation where both series are 0, and does not prove metric properties.

Contributions In this work we suggest a new formal definition of *State Transition Event timeseries* (STE-ts) and propose a *Selective Temporal Hamming* (STH) similarity and associated distance. STH has the following benefits:

- It is a formal generalization of the Hamming and Jaccard metrics for continuous time;
- It is equivalent to metrics computed on infinitely small sampling periods, avoiding any distortion;
- It generalizes the Jaccard metric to non-binary series, including ambiguous states ("excluded");
- It is a proper metric (incl. triangular inequality) in many cases and can thus be easily used in a kernel;
- Its linear complexity $\mathcal{O}(n+m)$ solely depends on the number of events, as opposed to resampled Hamming and Jaccard $\mathcal{O}\big(n+m+n^{(P)}\big)$ that also depend on the number $n^{(P)}$ of sampling buckets.

The rest of this paper is organized as follows: in Section 3 we remind DES, define state transition event timeseries (STE-ts) and remind existing metrics for uniformly sampled STE-ts. In Section 4 we introduce novel distance metrics and detail a few interesting properties. In Section 5 we present experiments on simulated and real-world datasets, highlighting the benefits. Finally, we conclude in Section 6.

3. State Transition Event Timeseries

3.1. Discrete Event Systems

We consider a discrete event system **DES** with set of states $\mathcal{S} = \{\sigma_k\}$ and set $\mathcal{T}$ of possible state transitions defined as a set of pairs (before state, after state): $\mathcal{T} = \{tr_i = (b_i, a_i)\} \subset \mathcal{S} \times \mathcal{S}$. In the following work we focus on *state-changing transitions*, i.e. $b_i \neq a_i\ \forall i$. Note that any system whose state is represented by a finite number of categorical variables can be represented as such. This includes Markov Processes (a.k.a. Markov Chains) and HMMs. Fig. 1 illustrates two prototypic DES: a simplified DVD player system and an industrial alarm system with shelving capability. Note: *shelving* (a.k.a. *muting*) means that the true alarm state is unknown and not relevant according to the user, for example during a maintenance operation.

Property: (*Simplification*) merging some states and associated transitions. e.g., merging "DVD playing" and "DVD paused", leads to valid DES representations.

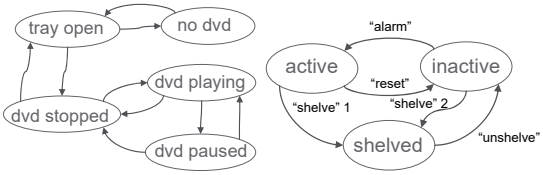


Fig. 1. Simple DVD Player (left); Alarm system (right).

3.2. State Transition Event Sequences and Timeseries

We define a *sequence* of state transition events **STE-seq** of length n as an ordered collection of n state-changing transitions $\{tr_k = (b_k, a_k)\}_{k=1,\dots,n} \subset \mathcal{T}^n$, with state-changing constraint $\{b_k \neq a_k\}\ \forall k$ and sequence consistency constraint $\{a_k = b_{k+1}\}\ \forall k < n$. Such a sequence can be rewritten as a *categorical sequence* of $n+1$ states $\{\sigma_k\}_{k=0\dots,n} \subset \mathcal{S}^{n+1}$ with state changing constraint $\{\sigma_k \neq \sigma_{k+1}\}\ \forall k < n$.

Let Ω be the set of all timestamps. We define a *timeseries* of state transition events **STE-ts** of length n, as an ordered collection of n timestamped state-changing transitions $\{(t_k, tr_k)\}_{k=1,\dots,n} \subset (\Omega \times \mathcal{T})^n$ complying with the above constraints, associated with a *start time* t_0 and an *end time* t_{n+1}. Such a timeseries can be rewritten as a *categorical timeseries* with $n+1$ pairs of state value and timestamp $\{(t_k, \sigma_k)\}_{k=0,\dots,n} \subset (\Omega \times \mathcal{S})^{n+1}$, associated with an end time. A given state σ_k is valid for the interval $[t_k, t_{k+1}[$.

Note that the above definition is similar to the traditional definition of *categorical timeseries* (CTS) [16], with three key restrictions: (a) a state value is valid for the interval following the timestamp, until next timestamp; (b) there is an explicit end time; and (c) all values differ from previous.

3.3. Prior Art – Reference Resampled Metrics

We now consider two STE-ts s_i and s_j with n and m transition events respectively, with the same start and end time and total duration T. We remind below the two simplest, most commonly used metrics, defined on resampled series. Resampled series $s_i^{(P)}, s_j^{(P)}$ are derived with sampling period P, sampling rate $F = 1/P$, resulting in (finite) number of samples $n^{(P)} = T/P$. We note $\sigma_{ik}^{(P)}$ and $\sigma_{jk}^{(P)}$ the respective resampled values $\forall k \in 1..n^{(P)}$. The *Hamming* distance HD is obtained as a count of non-matching samples:

$$HD\left(s_i^{(P)}, s_j^{(P)}\right) = \left|k = 1..n^{(P)} \mid \sigma_{ik}^{(P)} \neq \sigma_{jk}^{(P)}\right| \qquad (1)$$

The *normalized Hamming* distance nHD is defined as $nHD = HD/n^{(P)}$. It represents the approximate proportion of the time during which the two series' values differ. The quality of the approximation globally increases with the sampling rate but is subject to local distortion (see Section 5.1). *HD* and *nHD* are proper metrics complying with *positivity*, *symmetry*, *distinguishability*, and *triangular inequality* [34]. We note H = *n-HD*; nH = *1-nHD* the associated similarities.

The Jaccard similarity index J is defined for a binary representation of feature presence/absence (2) on static observations [23]. It is also used in practice on resampled series as a biased alternative to Hamming where value "0" has less interest than "1" [18, 24].

$$J\left(s_i^{(P)}, s_j^{(P)}\right) = \frac{\left|k|\ \sigma_{ik}^{(P)} = \sigma_{jk}^{(P)} = 1\right|}{\left|k|\ \sigma_{ik}^{(P)} = \sigma_{jk}^{(P)} = 1\right| + \left|k|\ \sigma_{ik}^{(P)} = 1, \sigma_{jk}^{(P)} = 0\right| + \left|k|\sigma_{ik}^{(P)} = 0, \sigma_{jk}^{(P)} = 1\right|} \quad (2)$$

The Jaccard distance is defined as $JD = 1 - J$ with $J(\mathbf{0}, \mathbf{0}) := 1$. It meets positivity, symmetry, distinguishability, and triangular inequality [23, 35].

Property: (*Linear complexity*) Assuming initial ordering, resampling of a single timeseries s_i is $\mathcal{O}\left(n^{(P)} + n\right)$ as for each bucket there is a need to find the events that fall in this bucket. Therefore *H, HD, nH, nHD*, *J,* and *JD* have complexity $\mathcal{O}\left(n^{(P)} + n + m\right)$.

Note: in this paper we use the simplified notations (e.g. HD, JD) for all resampled metrics except for places where it matters to remind the *P*: $HD^{(P)}, JD^{(P)}$.

4. Temporal Metrics for STE-ts

We now consider the collection of time intervals $\mathcal{I}_{ij} = \{\iota_k = [t_k, t_{k+1}[\}$ defined by the union of all timestamps in s_i, s_j. Its cardinal depends on the number of timestamps that are common to s_i and s_j: $\max(n, m) + 2 \leq \left|\mathcal{I}_{ij}\right| \leq n + m + 2$. For each interval ι we define Δ_ι its duration, and σ_i^ι and σ_j^ι the respective states of s_i and s_j on ι. Note that $T = \sum_{\iota \in \mathcal{I}_{ij}} \Delta_\iota$.

4.1. Temporal Hamming Similarity and Distance

We define the *Temporal Hamming* Similarity *TH* as the sum of durations of all intervals in $\mathcal{I}_{ij}$ where the two series s_i, s_j have the same value (3).

$$TH\left(s_i, s_j\right) = \sum_{\iota \in \mathcal{I}_{ij} \mid \sigma_i^\iota = \sigma_j^\iota} \Delta_\iota \quad (3)$$

The *Normalized Temporal Hamming* similarity *nTH* is obtained by dividing *TH* by the total duration:

$$nTH\left(s_i, s_j\right) = \frac{TH\left(s_i, s_j\right)}{T} = \frac{\sum_{\iota \in \mathcal{I}_{ij} \mid \sigma_i^\iota = \sigma_j^\iota} \Delta_\iota}{\sum_{\iota \in \mathcal{I}_{ij}} \Delta_\iota} \quad (4)$$

We define for *TH* and *nTH* their associated distances $THD = T - TH$ and $nTHD = 1 - nTH$.

Property: (*Metric*) *THD* and *nTHD* are proper *distance* metrics. *TH* and *nTH* are *similarity* measures.

Property: (*nH equivalence*) when all events have uniform timestamps, all interval durations are identical. Temporal Hamming metrics equal 'resampled' Hamming: $nTH = nH, nTHD = nHD$.

Property: (*Limit*) *nTH* (resp. *nTHD*) is the limit of $nH^{(P)}$ (resp. $nHD^{(P)}$) as sampling period P approaches zero:

$$nTH[D]\left(s_i, s_j\right) = \lim_{P \to 0} nH[D]\left(s_i^{(P)}, s_j^{(P)}\right) \quad (5)$$

We similarly define Temporal Jaccard *TJ* for binary series (6), associated distance $TJD = 1 - TJ$, and note that analogous properties (J equivalence, sampling limit) hold.

$$TJ\left(s_i, s_j\right) = \frac{\sum_{\iota \in \mathcal{I}_{ij} \mid \sigma_i^\iota = \sigma_j^\iota = 1} \Delta_\iota}{T - \sum_{\iota \in \mathcal{I}_{ij} \mid \sigma_i^\iota = \sigma_j^\iota = 0} \Delta_\iota} \quad (6)$$

Property: (*T-S equivalence*) Temporal Jaccard *TJ* equals the *temporal similarity* as defined in [33] for single alarm sets $A = \{s_i\}, B = \{s_j\}$.

4.2. Selective Temporal Hamming metric

We now consider a state partition: $\mathcal{S} = \mathcal{S}_I \cup \mathcal{S}_O \cup \mathcal{S}_E$ with

- $\mathcal{S}_I$ a set of states of *interest*. For example "active", or {"DVD playing", "DVD paused"}. We define sim a similarity function between states in $\mathcal{S}_I$ as in [31];
- $\mathcal{S}_O$ a set of *other* states, e.g. "inactive", "DVD stopped";
- $\mathcal{S}_E$ a set of *excluded* (or "ambiguous") states. For example "shelved", or {"tray open", "no DVD"}.

We define the *Selective Temporal Hamming* (STH) similarity for two STE-ts s_i and s_j, for state sets $\{\mathcal{S}_I, \mathcal{S}_O\}$, as in (7):

$$STH_{\{\mathcal{S}_I, \mathcal{S}_O\}}\left(s_i, s_j\right) = \begin{cases} ?\ (undef) & if\ \forall \iota,\ \sigma_i^\iota \in \mathcal{S}_E\ or\ \sigma_i^\iota \in \mathcal{S}_E \\ 1 & if\ \forall \iota,\ \sigma_i^\iota \notin \mathcal{S}_I\ and\ \sigma_i^\iota \notin \mathcal{S}_I, \\ x\left(s_i, s_j\right) & otherwise \end{cases} \quad (7)$$

with

$$x\left(s_i, s_j\right) = \frac{\sum_{\iota \in \mathcal{I} \mid \sigma_i^\iota, \sigma_j^\iota \in \mathcal{S}_I^2} sim(\sigma_i^\iota, \sigma_j^\iota) \cdot \Delta_\iota}{\sum_{\iota \in \mathcal{I} \mid \sigma_i^\iota, \sigma_j^\iota \in \left(\mathcal{S}_I^2 \cup (\mathcal{S}_I \times \mathcal{S}_O) \cup (\mathcal{S}_O \times \mathcal{S}_I)\right)} \Delta_\iota} \quad (8)$$

In this paper we restrict our analysis to sim being the identity function: 1 when $\sigma_i^t = \sigma_j^t$ and 0 otherwise. The *STH* similarity can be interpreted as the ratio between the time during which both series have the same state *of interest* and the time during which at least one of the series has a state of interest. The *excluded* state discards intervals. We define the Selective Temporal Hamming *distance* as $STHD = 1 - STH$.

Property: (*Normalized*) STH values are in [0,1]. STH = 1 means that on all intervals, either both series have a state of interest and it is the same ($\sigma_i^t = \sigma_j^t$), or none of them has a state of interest. STH = 0 means that there is no interval during which s_i and s_j have the same value and this value is of interest.

Property: (*Definition*) STH is *undefined* when for each interval $\iota \in \mathcal{I}$, at least one series has a state in $\mathcal{S}_E$. This situation is similar to the one caused by missing data in real-valued timeseries. A good fallback value in this case is application dependent. The authors suggest zero (0) as default.

Property: (Hamming equivalence) STH with all states "of interest" ($\mathcal{S}_I = \mathcal{S}$) is the normalized Temporal Hamming nTH. As such, all properties of nTH hold.

Property: (*Jaccard equivalence*) for binary STE-ts, *STH* with $\mathcal{S}_I = \{1\}, \mathcal{S}_O = \{0\}, \mathcal{S}_E = \emptyset$ is the *Temporal Jaccard* TJ. As such, all properties of TJ hold.

Property: (*Metric*) When $\mathcal{S}_E = \emptyset$ and $|\mathcal{S}_O| \leq 1$, the *STHD* distance satisfies both the *positivity*, *symmetry*. *distinguishability*, and *triangular inequality*. $|\mathcal{S}_O| \leq 1$ can be relaxed if *distinguishability* is not needed. A proof is provided in [36].

Property: (*Complexity*) Both $[n]TH$, TJ and *STH* have a linear complexity with respect to the number of intervals $|\mathcal{I}_{ij}|$ induced by s_i and s_j, so are $\mathcal{O}(n+m)$.

Table 1 summarizes the various metrics and their properties, and the two state-of-the-art metrics without resampling, FTH [31] and TS [33], as reference.

Table 1. Metrics properties, along with speed and precision experiments summary. '*' means 'under conditions'.

method	binary	categ.	metric	complexity	speed	distortion
(resampled) Jaccard	x		x	$\mathcal{O}(n+m+\frac{T}{P})$	baseline	$0-100\%$ (Fig.4 right)
(resampled) Hamming	x	x	x			
Selective Temporal Hamming STH	x	x	*	$\mathcal{O}(n+m)$	$3.5-4950\times$ faster (Fig.2, Fig.4 left)	None
Temporal Hamming $= STH_{\mathcal{S},\emptyset}$	x	x	x			
Temporal Jaccard $= STH_{\{1\},\{0\}} = TS$	x		x			
Fuzzy Temporal Hamming FTH	x	x	?	$\mathcal{O}(max(n,m)^2 log(max(n,m)))$	N/A	N/A

5. Experiments

5.1. Execution Time and Precision Benchmarks

In this section we perform controlled experiments with simulated STE-ts to validate the two main benefits of *STH* over resampled metrics such as Hamming or Jaccard: execution time and precision. For value comparison purposes we use the Hamming configuration ($\mathcal{S}_I = \mathcal{S}$) for *STH*, so *STH* = *nTH* in this section and its values are compared to the resampled normalized Hamming $nH^{(P)}$. Yet, similar results can be observed by choosing a Jaccard configuration ($\mathcal{S}_I = 1, \mathcal{S}_O = 0$) and comparing *STH(= TJ)* to resampled $J^{(P)}$.

Results are obtained on a PC with an AMD Ryzen 5 pro 5650U 2.3 GHz processor and 16 Gb ram. Resampling is done with the `pandas` library [37, 38].

Execution time vs. number of events

We create *Rts(n)* a timeseries with *n* randomly occurring binary state change events spanning 30 days. We also create *Tk*, a timeseries spanning 30 days with 8640 evenly spaced (/5mins) events, with added random uniform time lags ranging from 1ms to 200 ms.

Fig. 2 (left) shows average computation times (over 30 runs) of *STH* and $nH^{(P)}$ with 5 min resampling, for pairs *Rts(n)* vs. *Rts(n)*, *Rts(n)* vs. *Rts(1)* and *Rts(n)* vs. *Tk*. Both metrics execution time grow linearly with the number of events in the series, confirming our complexity analysis. Fig. 2 (right) shows the execution time ratio t($nH^{(P)}$)/t(*STH*). *STH* is between 3.5-14 times faster than $nH^{(P)}$ depending on the series. The largest improvements [7.5x-14x] are obtained when one of the series has only one event (*Rts(1)*).

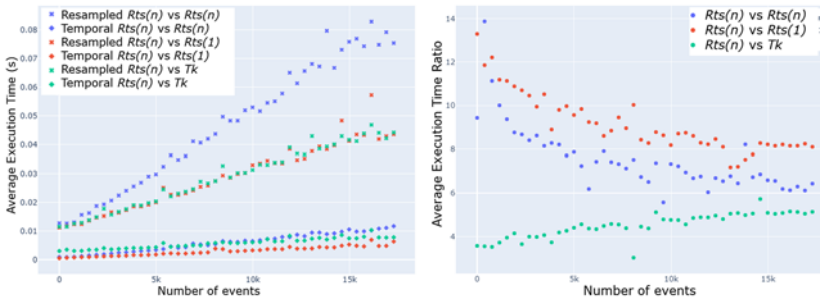


Fig. 2. Execution time (left) and ratio (right) for Resampled vs. Temporal Hamming distances for various pairs of series.

Execution time vs resampling period

We now generate two periodic binary STE-ts spanning 1 month (Fig. 3): PS_0 with 14min period and 60 % duty cycle; and $PS_{1/3}$ obtained by adding to PS_0 a lag of 1/3 of a period.

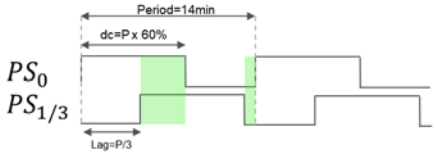


Fig. 3. Two periodic STE-ts with matching states highlighted.

We measure the average computation time over 30 runs of $nH(PS_0, PS_{1/3})$, for a wide range of resampling periods, and compare it with that of $STH = nTH$. Results in Fig. 4 (left) confirm that $nH^{(P)}$ is $O(T/P)$: as the sampling period P decreases, it is up to 4950× slower.

Metric precision

By design, PS_0 and $PS_{1/3}$ have the same state 1/3 of the time. Therefore, their normalized Hamming distance should be 2/3 if measured on an infinite number of periods. $STHD = nTHD$ is computed and is 0.6666574 as expected. We compute $nHD^{(P)}$ for various sampling periods P and compare it with $STHD$. Fig. 4 (right) shows the impact of resampling distortion on $nHD^{(P)}$. For some sampling periods the series are even considered identical by nHD ($nHD = 0$)! $STHD$ always guarantees the most accurate value.

Table 1 summarizes the results of this section.

5.2. Experiments on New Analysis Capabilities

In this section we illustrate the capability of STH to generalize *Jaccard* to non-binary series. We will use Clustering to highlight the impact of various choices of states to include in $\mathcal{S}_I$ and $\mathcal{S}_E$. For two public datasets below, raw data is first transformed to event transitions timeseries. We then compute a pairwise distance matrix for each pair of STE-ts, for various distance metrics. We finally apply agglomerative clustering on resulting distance matrices and comment.

Leveraging $\mathcal{S}_I$ to focus on specific states

We can focus STH on specific states of interest $\mathcal{S}_I$, the same way we use Jaccard to focus on the "1" in binary series. We illustrate this on the *US Weather Events* dataset, employed in a study to discover propagation and influential patterns [39]. It contains a collection of weather events data across 2071 US locations. Possible states are: *Cold, Fog, Snow, Hail, Rain, Storm, Precipitation,* and *Normal*, which represents the absence of notable weather event. Our analysis focuses on 2019, for 1000 random locations.

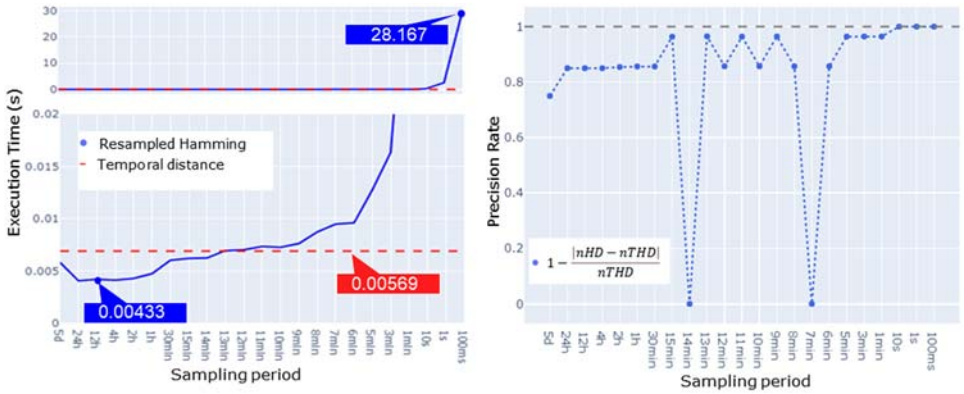


Fig. 4. Left: execution time (s) for resampled Hamming distance (blue) vs. STH=nTH (dashed red), for various sampling periods (x-axis). Bottom left: zoom on y-axis. Right: precision rate of nHD (1=no distortion) for various sampling periods.

We run clustering with 3 settings of STH. The number of clusters k is selected from the dendrogram based on decreasing the optimal number found by Silhouette score until an acceptable macroscopic geographical view is found. Settings are:

a) $\mathcal{S}_I = \mathcal{S}$ so $STH = nTH$ (Hamming); $k = 30$;
b) $\mathcal{S}_I = \mathcal{S} \backslash \{Normal\}$, $\mathcal{S}_O = \{Normal\}$; $k = 30$;
c) $\mathcal{S}_I = \{Snow, Hail\}, \mathcal{S}_O = \mathcal{S} \backslash \{Snow, Hail\}$; $k = 30$.

Results are shown in Fig. 5, where each marker in the map represents a location, and its color and shape represent the cluster id. Whereas Hamming (a) reveals a large-size 'blue' cluster, in (b) since the *Normal* (no event) state is removed from $\mathcal{S}_I$, differences between abnormal states are highlighted, for example on the US east coast; in (c) only snow and hail are in $\mathcal{S}_I$, highlighting differences in north-eastern regions but not so much in southern ones anymore.

Leveraging $\mathcal{S}_E$ to handle ambiguity

The new set $\mathcal{S}_E$ can be used to deal with ambiguous states where it is preferrable to not derive any similarity value. We illustrate this on the *MASS SS3 Sleep Annotations* dataset, that contains sleep stages annotations for an entire sleep period of 62 patients

[40]. The states are {?, *W,1,2,3,R}*. R (rapid eyes movement) is the state of interest, while W (awake), and the three levels of sleep depth (1, 2, 3) are not. The absence of annotation on an interval is replaced with a new *Sleep stage "E"* state.

We compare two STH settings:

a) $\mathcal{S}_I = \{R\}$ and $\mathcal{S}_E = \emptyset$ so *STH* is equivalent to *Jaccard* on a preprocessed dataset where R is mapped to 1 and all other states are mapped to 0;

b) $\mathcal{S}_I = \{R\}, \mathcal{S}_O = \{W, 1,2,3\}$ and $\mathcal{S}_E = \{?, E\}$.

Resulting similarities are then used in the agglomerative clustering algorithm, with a fixed number of clusters (8). The patients' STE-ts are displayed in the heatmap of Fig. 6 in the agglomeration order. Clusters are delimited by white stripes. We observe that clusters in (b) seem purer in terms of content, and that some individuals with many missing or ambiguous values are better grouped. For example, the individual identified by the arrow was isolated in a single cluster with Jaccard, and is now in a consistent group.

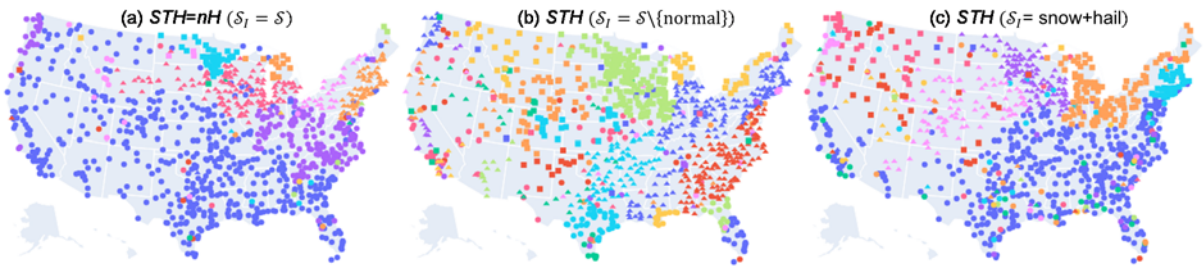


Fig. 5. US map with one marker per location, STH-based clusters identified by color and shape, for 3 sets of states of interest.

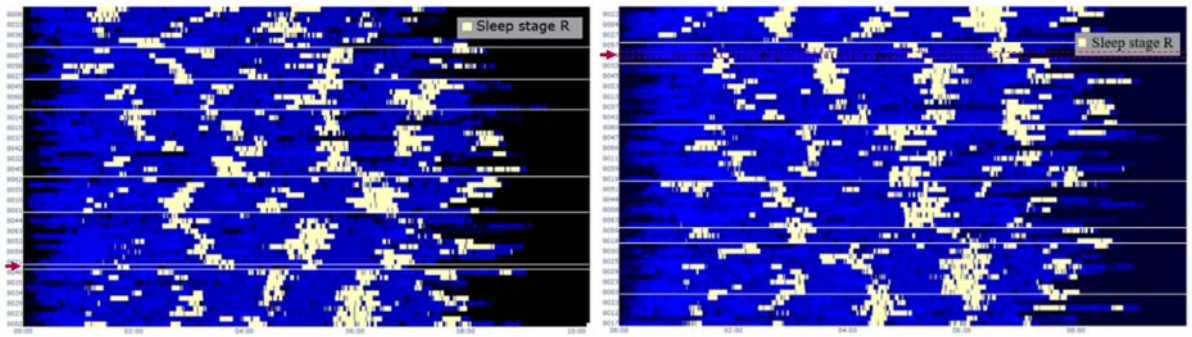


Fig. 6. Patients' (y axis) sleep state across time (x axis) sorted by aggregation order with 8 clusters (white stripes). Left: Jaccard {R vs. all}; right: $\mathcal{S}_I = \{R\}$, $\mathcal{S}_E\{?, E\}$ (both black) $\mathcal{S}_O = \{W, 1,2,3\}$ (four shades of blue).

6. Conclusion and Future Work

In this paper we formalized state transition event timeseries (STE-ts), bridging *event sequences* and *categorical timeseries* formalisms in the context of Discrete Event Systems. We introduced the *Selective Temporal Hamming* metric, generalizing *Hamming* and *Jaccard* for continuous time. As opposed to these resampled metrics where a tradeoff between speed and precision is required, our experiments on simulated datasets confirm that *STH* avoids distortion and is faster to compute, making it particularly suitable for large scale data analysis. Moreover, *STH* also brings *Jaccard*-like capabilities for non-binary series. Our clustering experiments on real world datasets highlighted the ability to focus on states of interest, and to handle ambiguous states – both key features to inject subject matter expertise in the analysis. These properties make *STH* particularly well suited to compare STE-ts while not limiting its applicability to any categorical timeseries.

In the future we plan to study how *STH* can be used in a kernel, e.g. for SVM classification tasks. Also, in this paper we restricted the state similarity weights *sim* to the identity matrix; it would be interesting to study the conditions on *sim* under which metric properties of *STH* still hold. Finally, finding a way to tolerate time shifts without impacting complexity significantly is another interesting challenge.

References

[1]. G. Haddeler, The analysis of discrete-event system in autonomous package delivery using legged robot and conveyor belt, *arXiv preprint*, 2021, arXiv:2101.12347.

[2]. P. J. G. Ramadge, W. M. Wonham, The control of discrete event systems, *Proceedings of the IEEE*, Vol. 77, Issue 1, 1989, pp. 81-98.

[3]. W. Hu, et al., An application of advanced alarm management tools to an oil sand extraction plant, *IFAC-PapersOnLine*, Vol. 48, Issue 8, 2015, pp. 641-646.

[4]. C.-H. Ng, S. Boon-Hee, Queueing Modelling Fundamentals: With Applications in Communication Networks, 2nd Ed., *Wiley*, 2008.

[5]. J. Zhao, Y. L. Chen, et al., Modeling and control of discrete event systems using finite state machines with variables and their applications in power grids, *Systems & Control Letters*, Vol. 61, Issue 1, 2012, pp. 212-222.

[6]. D. J. N. J. Soemers, et al., Spatial state-action features for general games, *Artificial Intelligence*, Vol. 321, 2023, 103937.

[7]. T. F. Liao, et al., Sequence analysis: its past, present, and future, *Social Science Research*, Vol. 107, 2022, 102772.

[8]. J. Opfer, K.-E. Gottschalk, Identifying discrete states of a biological system using a novel step detection algorithm, *PLoS ONE*, Vol. 7, Issue 11, 2012, e45896.

[9]. C. G. Cassandras, S. Lafortune, Introduction to Discrete Event Systems, 2nd Ed., *Springer*, 2008.

[10]. M. van de Velden, et al., A general framework for implementing distances for categorical variables, *Pattern Recognition*, Vol. 153, 2024, 110547.

[11]. M. J. Warrens, Similarity coefficients for binary data, PhD Thesis, *University of Leiden*, 2008.

[12]. I. Morlini, S. Zani, A new class of weighted similarity indices using polytomous variables, *Journal of Classification*, Vol. 29, 2012, pp. 199-226.

[13]. T. Rahier, S. Marié, F. Forbes, A pre-screening approach for faster Bayesian network structure learning, in Machine Learning and Knowledge Discovery in Databases. Research Track, *Springer*, 2023, pp. 211-227.

[14]. Z. Šulc, H. Řezanková, Comparison of similarity measures for categorical data in hierarchical clustering, *Journal of Classification*, Vol. 36, 2019, pp. 58-72.

[15]. M. Studer, G. Ritschard, What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures, *Journal of the Royal Statistical Society Series A: Statistics in Society*, Vol. 179, Issue 2, 2016, pp. 481-511.

[16]. Á. López-Oriona, J. A. Vilar, P. D'Urso, Hard and soft clustering of categorical time series based on two novel distances with an application to biological sequences, *Information Sciences*, Vol. 624, 2023, pp. 467-492.

[17]. S. Charbonnier, N. Bouchair, P. Gayet, Fault isolation by comparing alarm lists using a symbolic sequence matching algorithm, *IFAC Proceedings Volumes*, Vol. 47, Issue 3, 2014, pp. 7085-7090.

[18]. M. Lucke, M. Chioua, et al., Advances in alarm data analysis with a practical application to online alarm flood classification, *Journal of Process Control*, Vol. 79, 2019, pp. 56-71.

[19]. J. Wang, H. Li, J. Huang, C. Su, Association rules mining based analysis of consequential alarm sequences in chemical processes, *Journal of Loss Prevention in the Process Industries*, Vol. 41, 2016, pp. 178-185.

[20]. Y. Laumonier, J.-M. Faure, et al., Discovering systematic relations between alarms for alarm flows reduction, in *Proceedings of the 6th International Conference on Control, Decision and Information Technologies (CoDIT'19)*, 2019, pp. 1055-1060.

[21]. L. Lesnard, Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns, *Sociological Methods & Research*, Vol. 38, Issue 3, 2010, pp. 389-419.

[22]. S. Frühwirth-Schnatter, C. Pamminger, Model-based clustering of categorical time series, *Bayesian Analysis*, Vol. 5, Issue 2, 2010, pp. 345-368.

[23]. M.-J. Lesot, M. Rifqi, H. Benhadda, Similarity measures for binary and numerical data: a survey, *International Journal of Knowledge Engineering and Soft Data Paradigms*, Vol. 1, Issue 1, 2009, pp. 63-84.

[24]. S. R. Kondaveeti, et al., Graphical tools for routine assessment of industrial alarm systems, *Computers & Chemical Engineering*, Vol. 46, 2012, pp. 39-47.

[25]. A. Sadr, M. Zolfaghari-Nejad, Weighted Hamming distance for PUF performance evaluation, *Electronics Letters*, Vol. 49, 2013, pp. 1376-1378.

[26]. P. Montero, J. Vilar, TSclust: an R package for time series clustering, *Journal of Statistical Software*, Vol. 62, Issue 1, 2014, pp. 1-43.

[27]. C.-T. Do, et al., Multi-modal and multi-scale temporal metric learning for a robust time series nearest neighbors classification, *Information Sciences*, Vol. 418-419, 2017, pp. 497-513.

[28]. C. -C. M. Yeh, et al., Matrix Profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets, in *Proceedings of the IEEE 16th International Conference on Data Mining (ICDM'16)*, 2016, pp. 1317-1322.

[29]. P.-F. Marteau, Time warp edit distance with stiffness adjustment for time series matching, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, Issue 2, 2009, pp. 306-318.

[30]. S. de Waele, P. M. T. Broersen, Error measures for resampled irregular data, *IEEE Transactions on Instrumentation and Measurement*, Vol. 49, Issue 2, 2000, pp. 216-222.

[31]. C. Moreau, et al., A fuzzy generalisation of the Hamming distance for temporal sequences, in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'21)*, 2021, pp. 1-8.

[32]. Q. H. Liu, N. Nguyen, Nonuniform fast Fourier transform algorithm and its applications, in *Proceedings of the IEEE Antennas and Propagation Society International Symposium*, 1998.

[33]. H. Reinhardt, J. -P. Bergmann, et al., Temporal analysis of event-discrete alarm data for improved manufacturing, *Procedia CIRP*, Vol. 93, 2020, pp. 742-746.

[34]. M. M. Deza, E. Deza, Encyclopedia of Distances, 2nd Ed., *Springer*, 2013.

[35]. S. Kosub, A note on the triangle inequality for the Jaccard distance, *Pattern Recognition Letters*, Vol. 120, 2019, pp. 36-38.

[36]. S. Marié, Proof of metric properties for the selective temporal Hamming distance, HAL open archive preprint, 2025, https://hal.science/hal-05328460

[37]. The pandas development team, pandas-dev/pandas: Pandas, *Zenodo*, 2020, DOI: 10.5281/zenodo.3509134

[38]. W. McKinney, Data structures for statistical computing in Python, in *Proceedings of the 9th Python in Science Conference (SciPy'10)*, 2010, pp. 56-61.

[39]. S. Moosavi, et al., Short and long-term pattern discovery over large-scale geo-spatiotemporal data, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2727-2735.

[40]. CEAMS, SS3 Sleep Annotations, Borealis, the Canadian Dataverse Repository, V1, 2022. DOI: 10.5683/SP3/YD8AYI.

(020)

Above, On, and Below the Surface: Data Services in Large Collaborative Projects

V. Vassilev [1], G. Petkov [1], B. Kraychev [1], S. Haydushki [1], V. Sowinski-Mydlarz [2], S. Nikolov [1], N. Shivarov [1] and DiverSea Project Partners

[1] Sofia University, GATE Institute, 5, James Bourchier Blvd, 1164 Sofia, Bulgaria
[2] London Metropolitan University, 166-220 Holloway Road, N7 8DB London, UK
Tel.: + 359 878057265
E-mail: vassil.vassilev@gate-ai.eu

Summary: The collaborative projects under the HORIZON Europe framework programs of the European Union typically involve a large number of partners from different countries. The data-centric projects amongst them often require integration of multiple data sources, diverse data formats in different modes of data collection, which leads to complex data management architectures and policies. This article explores some of the design decisions, organisational principles, and technological solutions for addressing them by focusing on integration and hybridization. The research has been conducted while working on DiverSea, a project dedicated to the analysis of the biodiversity dynamics along the shores of Europe from the Black Sea, along the entire Mediterranean from East to West and all the way up to the North Sea, but the takeaways are important for many other collaborative projects, which face similar issues not only in the water, but also on the land and in the air.

Keywords: Data spaces, Data integration, Data storage and indexing, Visualization and data services, AI technologies.

1 Introduction

Many joint projects funded in the framework of Horizon Europe of the EU include significant information processing, which requires a proper data management policy. This article presents our experience in this direction in DiverSea [1], one of the three projects dedicated specifically to the biodiversity of the coastal seas around Europe, its ecosystem, and dynamics (the other two being Marco Bolo [2] and OBAMA-NEXT [3]). We believe it applies not only to the specific maritime research and innovation topics, but to many environmental projects in general as well.

In data-centric projects, data management includes multiple data processing operations that start with the collection of raw data and end in the interpretation of the consolidated, integrated and pre-processed data and its analysis. In DiverSea, for example, the focus is on analysis of the dynamics of sea biodiversity along the shores of Europe from the Black Sea in the South-East along the entire Mediterranean coast, all the way up to the North Sea in the North-West. The data in such projects typically comes from separate case studies, conducted by the different project teams with catchment areas in close proximity to the partner locations. Due to the big diversity of data granularity, formats, modes of collection, and communication protocols, it is becoming important to organize the data management policies around Big Data Technologies, even if the volume of data is not big. This leads to the necessity to dedicate adequate technical resources which in DiverSea project is isolated in three separate workpackages (see Fig. 1). GATE Institute of Sofia University is responsible for coordination of all data management operations (Workpackage 2) with additional role in the preparation of the data for analysis (workpackages WP3 and WP4).

2. Design Considerations

2.1. Distributed vs. Centralized Data Management

There are several different architectural solutions for Big Data Management, which lead to different organization of the data processing workflows and impose different requirements on the technical staff involved in data processing on each partner side.

Centralized Architecture: Based on full centralization of the analytical operations and rooted in the oldest *data warehousing* [4]. Heavily used in many business projects at the end of the last century, further elaborated in the *digital twins* [5] with many applications outside business. Light technical requirements for data providers, heavy expertise in the central coordinating site.

Distributed Architecture: Full decentralization of the data processing with an equal role for all data providers. Modern solution, which leads to the implementation of complex systems such as *data spaces* [6] and *blockchains* [7] require significant technical capacity and dedicated resources from each participant.

Hybrid Architecture: As a combination of the previous two, it has the advantage of being simpler than the distributed and more flexible than the centralized architectures, allowing distribution of data consumers' responsibilities while lowering the technical requirements of the data providers.

The solution implemented in DiverSea is hybrid. Although there are clear benefits for more modern, fully distributed architecture, especially for projects on Pan-European level, the leading role in such projects as a rule is taken by non-technical organisations, and this, unfortunately, shifts the focus towards the data generation while somewhat neglecting the data management and analysis using the recent statistical and machine learning methodologies. In the context of the sea biodiversity, for example, it would be of strategic advantage to have one project, which unites the three complementing projects – DiverSea, Obama-Next, and Marco-Bolo, and implement a common Data Space of European Sea and Ocean Biodiversity. This would allow combining the complementary both the data and the data services to implement more elaborate data management strategies, including more valuable data analytics pipelines.

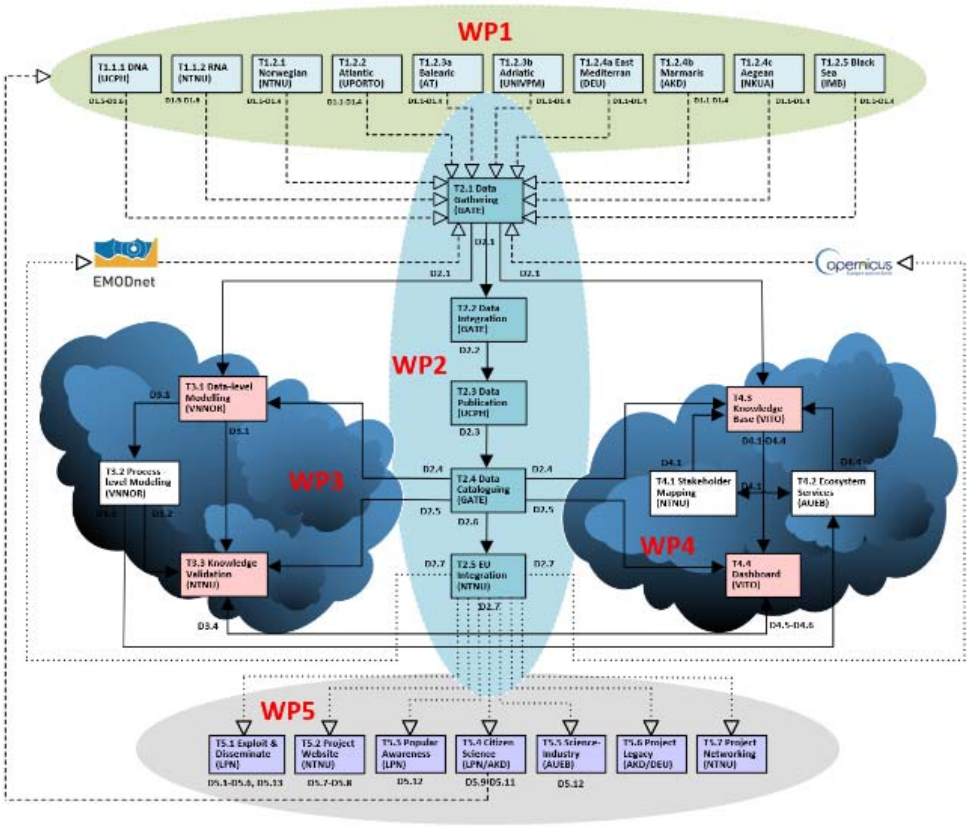


Fig. 1. Data Management Structure of DiverSea Project.

2.2. Data vs. Metadata

The data-centric projects start with gathering the raw data from the field studies. The most important artefact from data management point of view at this stage is the meta-data description, which is the entry point for developing the data models of the project data space. Following the common understanding that the meta-data is information about the data, however, it is often overlooked that the meta-data is not only about the structure and content of the physical data, but also a source of additional information to help building the data space architecture (see Table 1).

Depending on the nature of data as specified in the metadata the data model can differ substantially – it can be purely relational, object-relational or object-oriented, tree, graph or file-based. For well-structured datasets the most appropriate is the purely relational format, since it is easily represented in standard relational databases, but when the data comes from sensors, drones or satellites more suitable is the object-relational, object-oriented or graph format. Completely unstructured data, like images and videos coming from sound recordings, scanning or photographing are naturally kept in separate files, Finally, in the case of very large volume of data popular Big Data repositories such as Hadoop or Cassandra [8, 9] can be used instead. This data can be also complemented by external data, found in public repositories such as Copernicus or EmodNet.

2.3. Physical vs. Logical Model

Although the initial efforts triggering the data space development belong to the domain specialists collecting the data, some input from the data analysts can make the meta-data more informative. The

metadata can feed directly into the physical data model, but it can also help developing the domain ontology by providing information about the partitioning, taxonomic classification and dependencies across datasets. The ontological model of the data space allows to add two important additions to the data in it: conceptual abstraction, which can easily be modeled using taxonomies, and use cases for data processing, which can support the data processing and form a body of knowledge on its own on a conceptual level. A fragment of the ontological model of DiverSea presenting the most important EOV and EBV factors [10] of the analysis of the biodiversity is shown on Fig. 2.

Table 1. Some important meta-data characteristics.

Data	Data Sources	Ingestion	Transportation
Data samples (structured data)	IoT, computing devices & networks	One-off	memory sharing, parameter passing
Messages (semi-structured data)	Emails, Messages, Logs, Alerts	One-off	MQTT, AMQP, SMS, SOAP, etc.
Artefacts (unstructured data)	Documents, Images, Movies	One-off	FTP, HTTP
Datasets (structured collections)	Table exports, CSV files	One-off, Batch	FTP, HTTP
Streams (semi-structured collections)	Dynamic URLs, Service APIs	Continuous	native to the API
Data files (fully/semi/ unstructured collections)	Static URLs, File systems	One-off	FTP, HTTP, SCP, WebDAV, etc.
Repositories (collections of collections)	Databases, Data Marts & Data Lakes	Batch	native to the repository

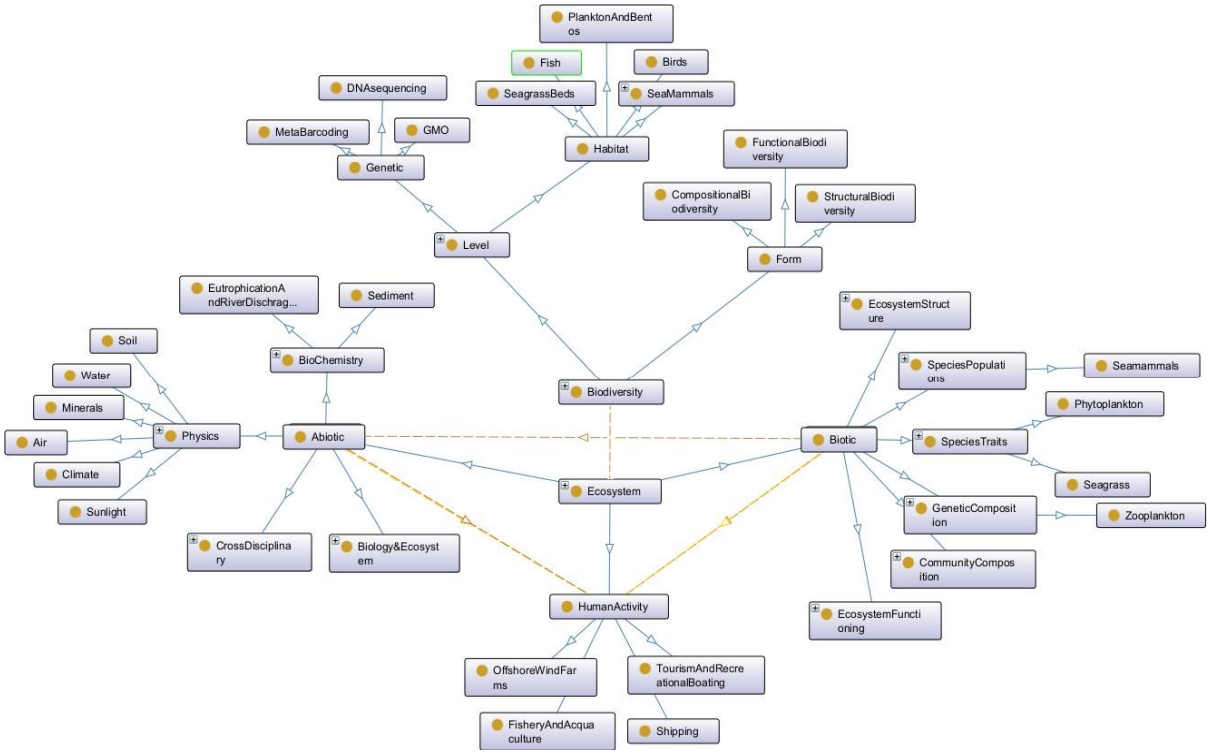


Fig. 2. Data ontology (fragment).

2.3. Data Management Tasks

The data space organization starts with the metadata obtained from the domain specialists, and the raw data generated within the field studies. The raw data is archived in its original format. Subsequently it is pre-processed and stored in suitable data repositories. The metadata is used also to create the data ontology which in addition to content information and logical structure of the data represents information about the datasets and the way they have been collected and transmitted (Fig. 3). This information is important for both the architecture of the data space and the subsequent data analysis. It is also used for some of the data services supported within the data space of the project.

3. DiverSea Data Space

In this section we will present the core of the data space as implemented in DiverSea project.

3.1. Data Model

The data model of DiverSea data space is hybrid. The structured data is modelled as purely relational (Fig. 4) and is stored in a separate relational database for each separate case study (https://datasets-diversea.gate-ai.eu/). On the other hand, the large files contain sound recordings from underwater drones and photogrammetric images from surface drones and satellites are stored directly in rge server file system

and are accessible over the Internet (https://datafiles-diversea.gate-ai.eu/).

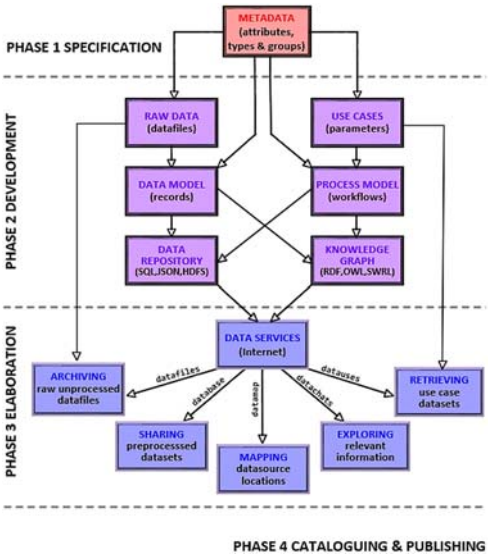


Fig. 3. Data Management Task Workflow.

The model is relatively small due to the strict focus on sea biodiversity and its dynamics. It consists of only 70 tables, structured into 6 groups according to different criteria for clustering the information – by content, technology, location, etc. We have decided to implement identical data models across all project case studies to enforce unification and support potential cross-location analysis. Although this creates some redundancy in the database implementations across the case studies, it also simplifies the maintenance.

3.2 Data Repositories

Since DiverSea project case studies generate only well-structured or completely unstructured data, we adopted PostgreSQL as a repository for storing the structured data after some pre-processing and Internet file system for storing the unstructured data in raw format. Both repositories are maintained centrally and have identical structure which simplifies the maintenance.

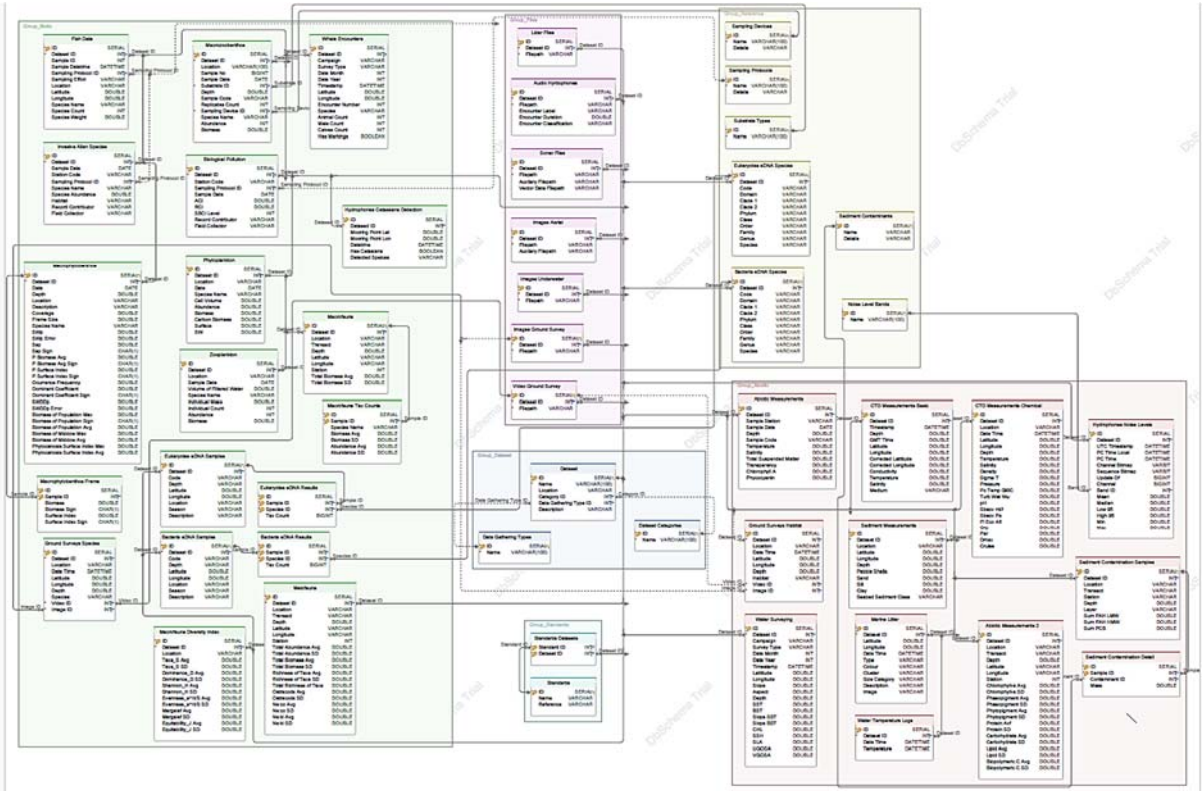


Fig. 4. Database for storing the pre-processed data.

3.3. Data Ontology

DiverSea data ontology is a graph model which provides higher level model of the concepts and relationships between them within the DiverSea data space. It supports not only the data sharing but also other data services convenient for the subsequent data analysis. It has been developed using Protégé and is currently stored as pure RDF graph in the graph database GraphDB.

3.4. Data Services

Although the data space is primarily concerned with collecting and sharing data, it can also provide

some useful data services. They incorporate expert knowledge from the problem domain, information about the environment, as well as technical information in supports of the data analysis. DiverSea data space currently supports five data services which allow quick look at the data, the environment in which the data has been collected and the context of its use.

Raw Data Archive

The raw data obtained from the project is kept in its original format in a file repository. It can be browsed from standard Web browser for inspection by the data providers, for initial review by the data consumers and auditing purposes. The repository is implemented as a protected file system, accessible over the Internet (Fig. 5). It is government by a system of priorities agreed by the data providers, but without write permission to protect the original data. At the end of the project some of this data will be published on open repositories such as EnmodNet and Copernicus for use by researchers.

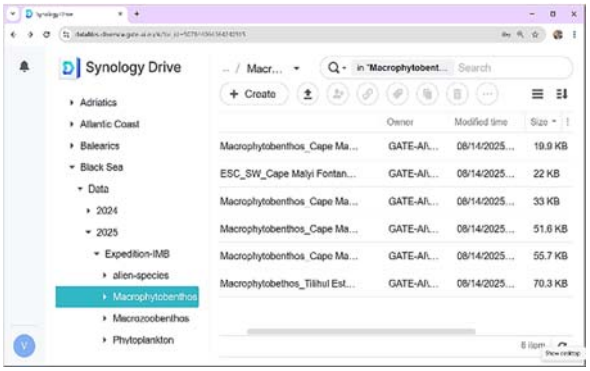


Fig. 5. Internet File System for Storing Raw Data.

Structured Data Repository

The raw data in DiverSea case studies does not grow very quickly due to the few periodic intakes of the fields teams, so no scalability issues were encountered by the data management team. Since the sensor and measurement data produced within the separate case studies is either in well-structured format (typical for EOV and EBV measurements), or in completely unstructured format (for various photogrammetric and scan data from cameras, drones and satellites), we combined the raw data file repository for unstructured data with a relational database for the structured data. Due to the relatively small amount of data we have chosen PostgreSQL DBMS as a database tool to accommodate the structured data after preliminary processing for filtering, completion and unification. The access to the databases of the case studies is granted to all project partners over the Internet through the PostgreSQL command interface in accordance with a flat profile model. The partners can search the data and export tables to support their analysis using standard SQL language (Fig. 6).

Small problem in this solution is the security issue created by the use of database console for accessing the data – the authorized users can enter the database over the Internet, which opens the door for potential malicious interventions through the DBMS system. This problem can be fixed by the use of VPN, but it would complicate the operations. The alternative is tio build a separate information system frontend with strict profile control, but it would be an overkill considering the relatively small amount of data and the limited project resources.

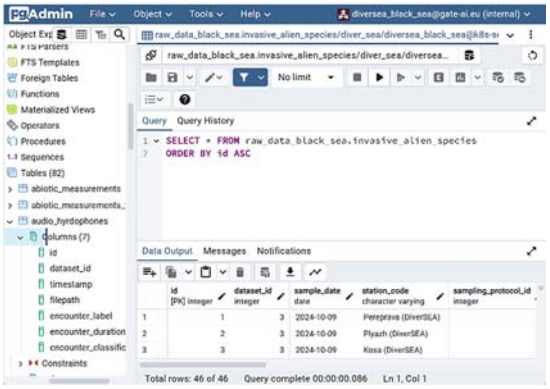


Fig. 6. Database Repository for Structured Data.

Mapping Data Sources

Geographic localization and mapping of the project data sources is important for both external presentation of the case studies and for internal preparation of the data analysis. We have experimented with several mapping services (ESRI, Cesium, Open Street Map) before selecting our final choice, TerriaJS (Fig. 7). It has some limitation in comparison with other services, but also advantages – it is compatible with many mapping formats, and provides free maps of the seas, including maps from EmodNet which are of particular interest for the project.

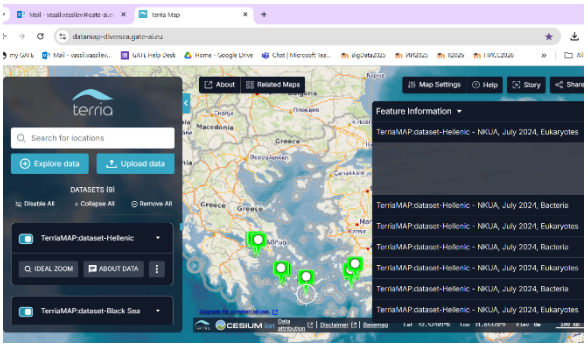


Fig. 7. Data Sources on the Map.

Looking for Answers Outside the Data Space

The increased competence of the LLM models provides new possibilities for expanding the data exploration and analysis beyond the limitation of the

command languages. Fig. 8 shows a session of interaction with DiverSea chatbot, which is integrated with the data space. The bot possesses a broad knowledge base about the sea, its marine life, and the project case studies and provides the end users with an immersive experience of the Earth's water basins. Our bot supports interaction in most European languages, this making it a valuable resource for learning and discovery of relevant information in an easy way.

Semantic Search

DiverSea data space stores data which can be analysed using different methods – correlation across locations, discovering historical trends, estimating impacts of the environment changes and human activities, etc. The physical level of the dataspace does not support adequately the search and retrieval of data relevant to these use cases due to the limited information in the metadata and the impossibility to establish cross dependencies between datasets. But more importantly, for successful analysis of the data it is required to have a good knowledge of the data, its location and relevance as well as intimate understanding of its internal representation within the numerous data repositories.

Fig. 8. Looking for Answers from LLMs.

Key to address this problem is the utilization of the data ontology for searching logically within the ontology instead of searching the physical data inside the databases. The semantic search service enables analysts to explore the data from the perspective of its relevance to the analytical use cases, bypassing the use of exploratory tools and the knowledge of protocols for interaction. This leads to a hybrid search approach in which the necessary data is retrieved after conceptual search within the data space ontology, which acts as a semantic index of the data.

To implement this concept we expanded the ontology with the most common use cases for data analysis and stored it in a dedicated graph database, GraphDB, which exposes SPARQL as a standard query language for searching the knowledge graphs. We have implemented a database plugin, which hides the command interface behind a graphical UI capable of translating the menu-driven request form into free format SPARQL queries and exeties them against the ontology (Fig. 9).

Currently our plugin allows to make queries focused on the datasets needed for executing analytical use cases, but we are continuing working on expanding the plugin with query templates for seeking specific descriptions of the parameters, such as location, time, specific descriptions and dependencies.

4. Current State and Future Development

The initial version of DiverSea data space has been completed and validated by the internal data providers and data consumers. All data in the data space is available to the project participants over the Internet in both raw and structured formats according to the agreed access profiles. Currently we are working on the second version of the data space, which incorporates the additional data services in support of the data analysis demonstrated here.

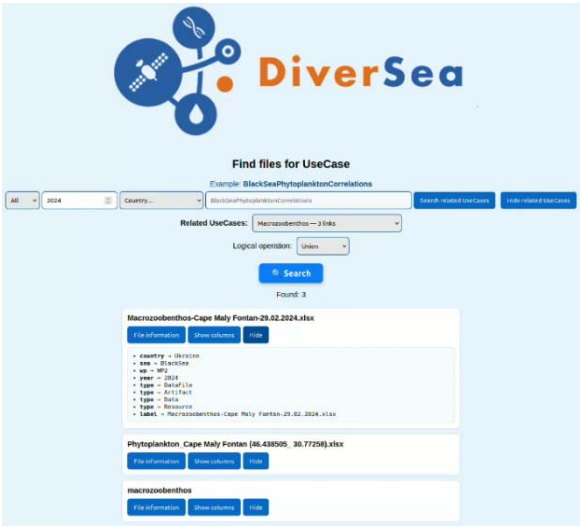


Fig. 9. Semantic Search for Datasets.

To make this experience really useful we are in conversation with our partners on the extension of the data space ontology with mini ontological models of the analytic use cases, so that we can search not only for datasets which are needed but also for direct and indirect links within the ontology. At the final phase of the project we will submit curated datasets to be published in the public repositories focused on the biodiversity such as Copernicus and EmodNet.

5. Conclusion

This article presents the experience of the DiverSea project data management team in developing data spaces within collaborative projects with multiple data providers and data consumers, real data collected in different formats and modes of operation. Although focused primarily on DiverSea, we believe that our methodology for data management, some of our ideas

for organizing the data spaces and very real experience are valuable for other data-centric collaborative projects.

Acknowledgements

The DiverSea project is funded by the European Union under the Horizon Europe Programme, Grant Agreement No. 101082004. Views and opinions expressed here are, however, those of the authors only and do not necessarily reflect those of the European Research Executive Agency (REA) nor the project partners.

References

[1]. DIVERSEA Project, Integrated observation, mapping, monitoring and prediction for functional biodiversity of coastal seas, https://www.ntnu.edu/diversea

[2]. MARCO-BOLO Project, Marine coastal biodiversity long-term observations: strengthening biodiversity observation in support of decision making, https://marcobolo-project.eu/

[3]. OBAMA-NEXT Project, Observing and mapping marine ecosystems – next generation tools, https://obama-next.eu/

[4]. R. Kimball, M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Ed., *Wiley*, 2013.

[5]. R. Khan, R. Young, Digital Twin & Digital Development's Handbook, *Independently published*, 2023.

[6]. International Data Spaces Association, Reference Architecture Model, GitHub Repository, https://github.com/International-Data-Spaces-Association/IDS-RAM_4_0/tree/main/documentation

[7]. I. Bashir, Mastering Blockchain: Inner workings of blockchain, from cryptography and decentralized identities, 4th Ed., *Packt Publishing*, 2023.

[8]. A. Anjomshoaa, et al., Data platforms for data spaces, in Data Spaces: Design, Deployment, and Future Directions (E. Curry, et al., Eds.), *Springer*, 2022, pp. 27-39.

[9]. V. Vassilev, V. Sowinski-Mydlarz, et al., Building a big data platform using software without license costs, in Open Source Horizons-Challenges and Opportunities for Collaboration and Innovation (L. Castro, Ed.), *IntechOpen*, 2024.

[10]. F. Muller-Karger, et al., Advancing marine biological observations and data requirements of the complementary essential ocean variables (EOVs) and essential biodiversity variables (EBVs) frameworks, *Frontiers in Marine Science*, Vol. 5, 2018, 211.

(021)

Quantifying the Unstructured Narrative of Patient Care in EHR Data

Edward Kim [1,2], Richard Foty [1] and Vicki Seyfert-Margolis [1]
[1] RespondHealth, Bethesda, MD, USA
[2] Department of Computer Science, Drexel University, Philadelphia, PA, USA
E-mail: ek826@drexel.edu, rich.foty@respondhealth.net, vicki.seyfert-margolis@respondhealth.net

Summary: Electronic Health Records (EHRs) contain vast amounts of patient information, yet much of their clinically relevant content is locked within unstructured narrative text. This case study applies a generative AI-based pipeline to extract, link, and verify medical concepts from unstructured clinical notes, evaluating their alignment with structured EHR data. Using over 110000 visits from more than 1000 practices, with a focus on 5000 Parkinson's disease patients, we quantify information volume, novelty, and consistency. Findings reveal significant redundancy, low novelty in longitudinal documentation, and persistent gaps between narrative and coded data. Our approach demonstrates how generative AI can surface high-value insights, improve structured and unstructured integration, and reduce the documentation burden in clinical practice.

Keywords: Generative AI, Named Entity Extraction, Electronic Health Records, Data Mining

1. Introduction

EHR systems are designed to provide a comprehensive and up-to-date account of a patient's medical history, supporting both clinical decision-making and research. In practice, however, these records are a heterogeneous mix of structured fields and unstructured narrative text, with the latter comprising roughly 80 % of the clinically relevant content [1]. While narrative documentation allows clinicians to capture nuanced details, it also leads to challenges in data extraction, clinician burden, redundancy, and misalignment with coded data. For example, clinicians now spend more time on the computer than with patients, i.e. nearly six hours of an eleven hour work day is consumed by EHR tasks, requiring an additional 1.4 hours of after clinic hour time [3]. The growing complexity of EHR content, coupled with the time burden placed on clinicians, underscores the need for advanced, scalable solutions that can efficiently transform unstructured narratives into actionable information.

Generative AI offers a promising path forward. Building on recent advances in large language models (LLMs), these systems can extract structured concepts, perform verification, and reconcile discrepancies across data modalities. This study develops and evaluates a generative AI pipeline to characterize information volume, identify novel content, and assess consistency between narrative notes and structured EHR fields.

2. Background

2.1. What is in an EHR Record?

EHRs capture a vast and massive array of patient data. Structured fields include demographics, ICD-coded diagnoses, problem lists, allergies, medications, vital signs, laboratory results, imaging, and procedure codes. However, approximately 80 % of EHR data is unstructured free text, such as clinical notes, imaging/procedure reports, and correspondence [5]. This imbalance means **that most clinical nuance and context reside in narrative form, which is not readily computable.** For example, a single patient in a pediatric ICU can generate over 1400 new data points per day, and a physician may be exposed to 16000-32000 data points in a single shift [6]. Large hospital EHRs may contain over 10000 unique medical codes and thousands of lab variables [7]. And despite the availability of these structured fields, clinicians often prefer narrative documentation for its flexibility and ability to capture details such as social determinants of health or diagnostic uncertainty [8]. Given this preference in documentation, a growing concern in EHR documentation is the prevalence of redundant content, often referred to as "note bloat". Modern EHRs facilitate copy-paste and auto-import, which can reduce typing but also lead to lengthy, repetitive notes. A UCSF study of 23000 inpatient notes found that **only 18 % of note text was newly authored, with 82 % copied or imported** [9]. Residents had the lowest proportion of new text (12 %), and more than half of their notes were copied forward. A 2022 analysis of 104 million notes at Penn Medicine found that 50 % of note text was duplicated, with duplication split between self and other authors [10]. Outpatient note length increased by 60 % from 2009 to 2018, and redundancy rose from 48 % to 59 % [4]. By 2018, only 29 % of outpatient note text was original. This heavy reuse can obscure new information, propagate errors, and make it difficult for clinicians to identify salient changes. The literature highlights a tension between efficiency (via reuse) and the risk of clutter and clinical error [10, 11, 5].

The informational **burden is amplified by discrepancies between what clinicians write and what they (or downstream coders) enter as discrete structured data**. Large-scale concept-coverage studies now show that only 13 % of clinical concepts

described in free-text have a corresponding structured code, and just 7 % of note-level concepts surface in structured form at the encounter level [2]. For example, a diagnosis may be mentioned in a note but not coded, or vice versa. Integrated workflows that link note writing and coding improve concordance (up to 88 % match), but perfect alignment is rare [11]. Essential details are often "hidden" in unstructured notes, especially when no appropriate code exists or when expressing uncertainty. Conversely, nearly one-quarter of medications listed in progress notes fail to reconcile with the formal EHR medication list, exposing patients to safety risks when decision-support logic relies solely on coded inputs [12].

2.2. How is AI Used to Improved EHR Documentation?

Given the dominance of unstructured text, natural language processing (NLP) and machine learning are essential for extracting clinical information from notes. Early systems used keyword and rule-based approaches, but recent advances leverage deep learning and transformer models (e.g., ClinicalBERT, GatorTron) trained on massive corpora of clinical text [1, 9]. Applications include computational phenotyping, adverse event detection, cohort identification, and automated chart review. For example, GatorTron was trained on 82 billion words of de-identified clinical notes and achieved state-of-the-art performance on multiple NLP tasks [8]. Integrating NLP-derived features with structured data improves predictive modeling and research [1]. New interventions such as medical scribes, voice recognition, team documentation, and novel note formats (e.g., APSO) have been explored to improve efficiency and satisfaction [13]. Integrated workflows, where coding and note writing are combined, improve accuracy and reduce redundancy [11]. However, challenges remain in accuracy, generalizability, interpretability, and privacy.

Traditional NLP approaches have attempted to bridge this gap, but often require intensive annotation and task-specific model training. Recent advances in Large Language Models (LLMs) and other generative AI provides an opportunity to address these limitations but require careful attention to data quality (especially LLM hallucinations), determinism of outputs, privacy, and interpretability. The relationship between narrative notes and structured data is an ongoing area of focus, where claims require support and alignment, yet are often lacking, with implications for both clinical care and research. Additionally, with the exponential improvement in generative AI towards generalizability, modular and trusted frameworks are needed to leverage these capabilities for clinical data extraction and verification [15, 16]. In this work, we present a generative AI-based pipeline to extract, link, and verify medical concepts from unstructured clinical notes, evaluating their alignment with structured EHR data. Using over 110000 visits from more than 1000 practices, with a focus on 5000 Parkinson's disease patients, we quantify information volume, novelty, and consistency of the data utilizing generative AI.

3. Methodology

In this case study, we utilized generative AI to process unstructured clinical notes from over 110000 encounters across three commercial EHR platforms. A Parkinson's disease cohort of 5000 patients was selected based on ICD-10 codes G20–G22, yielding 11250 encounters for detailed analysis. Our goal was three-fold. First, we wanted to quantify how computationally intensive it would be to use Large Language Models (LLMs) to temporally process, extract, and verify EHR data. Second, we wanted to quantify the amount of copy & paste in the record to see if we can automatically identify 'new' and relevant information. Third, our goal was to quantify the concordance (and discrepancy) between the coded data and unstructured clinical notes.

The first stage of the pipeline consisted of a structured JSON Extraction of clinical entities. We built a named entity recognition (NER) LLM extractor via prompting to identify key clinical entities, including medications, family history, disease status, adherence, phenotypes, and self-reported symptoms. Due to the fact that this is structured extraction, most off the shelf open source and cloud LLM APIs are usable [14]. There are two major issues around the use of LLMs for entity extraction. First are around the problems of trust and accuracy. We extensively address this issue in our previous work, see Kim et al. [15]. The second issue is around the efficiency of these methods, i.e. quantity, cost, and time. Here we quantify the total number of tokens that our processes take when extracting entities and validating these entities, see Fig. 1.

We note that these token totals processed are across all categories that we choose to extract and also account for the verification of these extractions via another LLM extraction loop. Thus, the total number of words in the EHR are significantly fewer, but the proportions and relative growth year over year are maintained. For the total breakdown of the tokens per category and verification, see Fig. 2. In our extraction, we ask for JSON categories such as the entity *name, status, is_new_information, additional_information, and atomic_statment* (for verification). Notably, the verification of extractions consume the greatest number of tokens, but are necessary for generative AI process. As noted by the relative increase in the size of the notes, there is a fairly significant jump in the past couple years, presumably around the time of the introduction of AI tools (like ambient listening and scribing tools); however, to confirm, more analysis would be required.

Information novelty was assessed longitudinally by comparing sequential visits for each patient, see Fig. 3. New information is quantified according to the entities

extracted at each visit that was not present in previous visits. For patients with 4 or more visits in the history, **our data shows that only 13 % of each clinical note is actually new.** New information is defined as entities that did not exist in previous visits. Entities that changed status are considered new information. While this does not directly indicate copy-paste forwarding, it is indicative and supports the research that shows that only 18 % of the clinical note is newly authored [9].

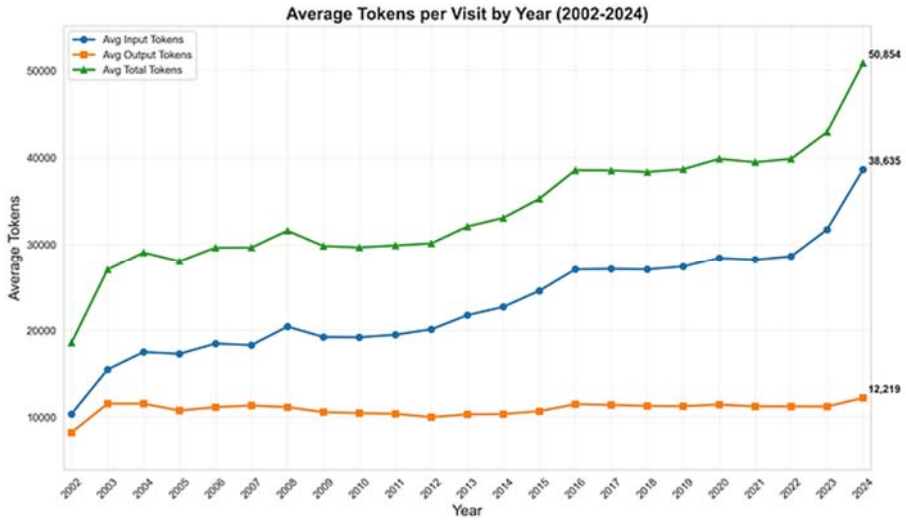


Fig. 1. Average input, output, and total tokens per clinical visit by year. The plot shows a steady increase in the average number of tokens per visit, with a marked acceleration after 2015 and again in 2023. This trend highlights the expanding volume of unstructured data in EHRs.

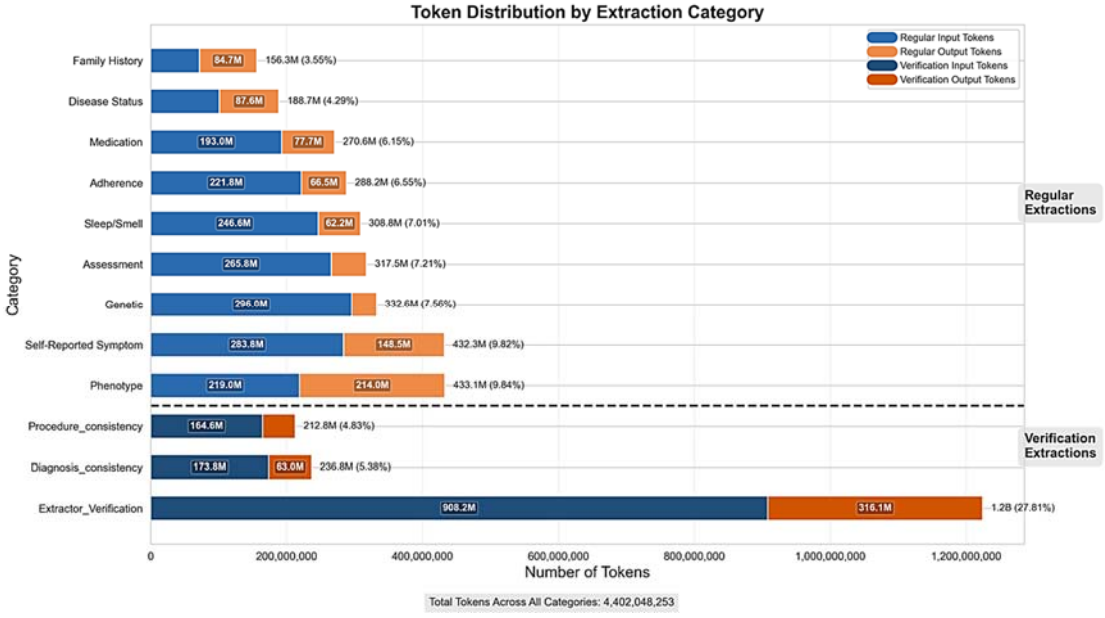


Fig. 2. Token distribution by extraction category. We break down the total number of input and output tokens processed by each extraction module, including both regular and verification extractions. Verification extractions account for the largest share of tokens, reflecting the resource intensity of consistency checks across diagnoses and procedures. The figure highlights the diversity of extraction tasks and the substantial computational load associated with large-scale EHR narrative analysis.

Finally, we built consistency verifiers and checks to compare the diagnoses and procedures described in the clinical note to the coded diagnosis and procedure data. The formal diagnoses in the clinical visit are captured by ICD10 codes and the procedures are captured by the CPT codes. We look at the diagnosis and procedure and run another LLM call that rates whether or not the code is, 3 = fully supported by the unstructured note, 2 = partially supported, 1 = potentially inferable but not partially or directly

mentioned, and 0 = not supported at all. We discovered low average verification scores for diagnoses (1.01 out of 3) and procedures (0.88 out of 3) indicating that, on average, most codes are only implied or partially supported by the clinical text, rather than being explicitly documented, see Table 3. Again, this supports the literature stating that only 13 % of clinical concepts described in free-text have a corresponding structured code [2].

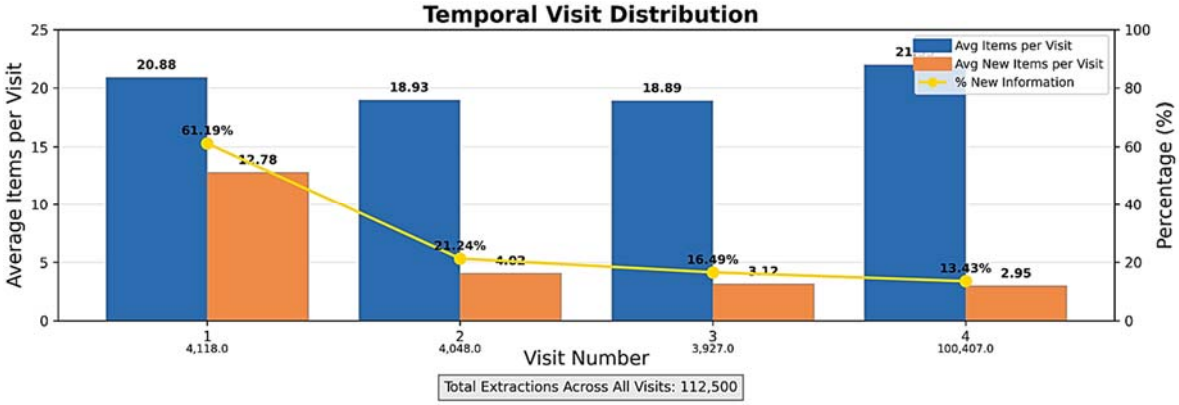


Fig. 3. Temporal visit distribution: Average items per visit, average new items per visit, and percentage of new information by visit number. Bar plots show the average number of extracted items (blue) and average number of new items (orange) for each visit, while the gold line indicates the percentage of new information contributed per visit. Extraction counts are annotated below each bar. The first visit yields the highest proportion of new information (61 %), with subsequent visits showing a marked decline in novelty (13– 21 %), reflecting the prevalence of redundant documentation and copy-forward practices in longitudinal EHR records.

Table 1. Consistency verification for diagnoses and procedures. "Total Items" is the number of extracted codes; "Total Items %" is the proportion of all extracted statements; "New Information Items" and "New Information %" indicate the number and proportion of codes representing new information; "Avg. Verification Score" is the average consistency score per code.

Category	Total Items	Total Items (%)	Avg Verification Score
Diagnoses	469446	19.24 %	1.01
Procedures	213472	8.75 %	0.88

4. Conclusions

This study reinforces that EHRs remain dominated by unstructured, often redundant content, with substantial gaps between narrative documentation and structured coding. Careful construction and utilization of generative AI offers a scalable solution to these challenges, enabling precise extraction, standardization, and verification of clinical concepts. Implementing such systems can reduce the documentation burden, improve structured–unstructured integration, and enhance the fidelity of EHR data for both clinical care and research.

As a primary goal of our work, we were able to quantify the content and concordance programmatically in EHR records. This groundwork enables us to focus on optimizing AI workflows to maximize the signal to noise ratio within the digital record and minimize clinician disruption, with the overall goal of enhancing patient care.

References

[1]. Z. Zeng, Y. Deng, X. Li, T. Naumann, et al., Natural language processing for EHR-based computational phenotyping, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 16, Issue 1, 2019, pp. 139-153.

[2]. T. Seinen, et al., Using structured codes and free-text notes to measure information complementarity in EHRs, *Journal of Medical Internet Research*, Vol. 27, 2025, e66910.

[3]. B. G. Arndt, J. W. Beasley, M. D. Watkinson, J. L. Temte, et al., Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations, *Annals of Family Medicine*, Vol. 15, Issue 5, 2017, pp. 419-426.

[4]. A. Rule, S. Bedrick, M. F. Chiang, M. R. Hribar, Length and redundancy of outpatient progress notes across a decade at an academic medical center, *JAMA Network Open*, Vol. 4, Issue 7, 2021, e2115334.

[5]. H.-J. Kong, Managing unstructured big data in healthcare system, *Healthcare Informatics Research*, Vol. 25, Issue 1, 2019, pp. 1-2.

[6]. L. Jalilian, S. Khairat, The next-generation electronic health record in the ICU: a focus on user-technology interface to optimize patient safety and quality, *Perspectives in Health Information Management*, Vol. 19, Issue 1, 2022, 1g.

[7]. B. Theodorou, C. Xiao, J. Sun, Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model, *Nature Communications*, Vol. 14, Issue 1, 2023, 5305.

[8]. X. Yang, A. Chen, N. PourNejatian, H. C. Shin, et al., A large language model for electronic health records, *NPJ Digital Medicine*, Vol. 5, Issue 1, 2022, 194.

[9]. M. D. Wang, R. Khanna, N. Najafi, Characterizing the source of text in electronic health record progress notes, *JAMA Internal Medicine*, Vol. 177, Issue 8, 2017, pp. 1212-1213.

[10]. J. Steinkamp, J. J. Kantrowitz, S. Airan-Javia, Prevalence and sources of duplicate information in the electronic medical record, *JAMA Network Open*, Vol. 5, Issue 9, 2022, e2233348.

[11]. T. S. Hwang, M. Thomas, M. Hribar, A. Chen, et al., The impact of documentation workflow on the accuracy of the coded diagnoses in the electronic health record, *Ophthalmology Science*, Vol. 4, Issue 1, 2024, 100409.

[12]. A. A. Monte, P. Anderson, J. A. Hoppe, R. M. Weinshilboum, et al., Accuracy of electronic medical record medication reconciliation in emergency department patients, *The Journal of Emergency Medicine*, Vol. 49, Issue 1, 2015, pp. 78-84.

[13]. A. J. Holmgren, C. A. Sinsky, L. Rotenstein, N. C. Apathy, National comparison of ambulatory physician electronic health record use across specialties, *Journal of General Internal Medicine*, Vol. 39, Issue 14, 2024, pp. 2868-2870.

[14]. E. Kim, M. Shrestha, R. Foty, T. DeLay, et al., Structured extraction of real world medical knowledge using LLMs for summarization and search, in *Proceedings of the IEEE International Conference on Big Data (BigData'24)*, 2024, pp. 3421-3430.

[15]. E. Kim, R. Foty, M. Shrestha, V. Seyfert-Margolis, Conformal prediction and verification of large language model extractions in EHR data, in *Proceedings of the SECURE-AI4H Symposium, AAAI Fall Symposium*, 2025.

(024)

Autonomous, Self-learning Cisco Digital Adoption Platform (CDAP) for Personalized, Proactive Campaign Creation and Targeted User Engagement

N. Kale
Cisco Systems Inc., 3700 Cisco Way, San Jose, CA 95134, USA
Tel.: +1 408 902 4609
E-mail: nikkal@cisco.com

Summary: This research introduces an Autonomous, Self-Learning Cisco Digital Adoption Platform (CDAP) that addresses critical limitations in traditional user engagement systems through advanced AI-driven personalization. Unlike conventional platforms relying on static segmentation, CDAP employs real-time behavioral analytics, federated learning with differential privacy ($\varepsilon = 1.0$), and sentiment-driven engagement optimization. The system implements a novel "cohort of one" approach using LSTM context encoding and Adam-optimized weight updates ($\eta = 1e^{-3}$), analyzing micro-interactions to optimize overlay placement and engagement timing. Through comprehensive experimental validation involving 66000+ participants across five industry domains over 8 weeks, CDAP demonstrates significant improvements: task completion rates increased from 72.3 % to 89.7 % ($p < 0.001$, Cohen's $d = 0.85$), user satisfaction improved by 40.3 %, time-to-proficiency reduced by 39.9 %, and support ticket volumes decreased by 148 %. The platform maintains less than 50ms inference latency while consuming 5627 CPU-hours total with federated coordination representing approximately 34 % of computational load, demonstrating practical scalability for enterprise deployment.

Keywords: Federated learning, Privacy-preserving analytics, Real-time personalization, Behavioral analytics, Sentiment analysis, Digital adoption, Cross-tenant optimization.

1. Introduction

The digital transformation landscape has witnessed unprecedented integration of complex enterprise software systems, creating substantial challenges for user adoption and proficiency development. Traditional Digital Adoption Platforms (DAPs) typically rely on static demographic segmentation and predefined content delivery models, failing to capture nuanced, real-time user interactions and contextual needs [1, 2]. This static approach results in delayed content delivery, irrelevant interventions, and suboptimal user experiences, leading to reduced engagement effectiveness and increased support burden [3].

Recent advances in machine learning and privacy-preserving technologies have opened new possibilities for intelligent, adaptive user engagement systems. Modern recommender systems demonstrate the effectiveness of personalized approaches at scale [4, 5], while federated learning frameworks enable collaborative learning across organizational boundaries while maintaining data sovereignty [6]. However, existing enterprise solutions have not successfully integrated these capabilities to address the specific challenges of digital adoption.

Contemporary research in user experience personalization emphasizes the importance of real-time behavioral analytics and sentiment-aware engagement optimization [7, 8]. Studies by Abbas et al. (2021) demonstrate significant improvements in user engagement through AI-driven behavioral analysis, while Kim & Kim (2021) show substantial benefits of personalization in digital banking environments [9]. However, these approaches have not been systematically applied to enterprise digital adoption with formal privacy guarantees.

The Cisco Digital Adoption Platform (CDAP) addresses these fundamental limitations through an autonomous, self-learning system that delivers personalized, proactive campaigns tailored to individual user behavior in real-time. This research presents comprehensive experimental validation demonstrating CDAP's superiority over traditional approaches across multiple performance dimensions through rigorous randomized controlled trial methodology.

2. Mathematical Framework

2.1. Adaptive Personalization Model

The core of CDAP's personalization system employs a sophisticated mathematical framework modeling user behavior, context, and historical patterns. The user behavior vector at time t captures multi-dimensional interaction signals:

$$\mathbf{U}(u,t) = \begin{bmatrix} \text{click\_rate}(u,t) \\ \text{dwell\_time}(u,t) \\ \text{scroll\_depth}(u,t) \\ \text{task\_completion}(u,t) \end{bmatrix}$$

Context encoding utilizes a 128-unit LSTM network to capture sequential interaction patterns:

$$\mathbf{C}(u,t) = \text{LSTM}(\text{sequence}, \mathbf{h}_{t-1}, \mathbf{c}_{t-1})$$

The adaptive personalization score integrates multiple components through learned weights:

$$P(u,t) = w_1\mathbf{U}(u,t) + w_2\mathbf{C}(u,t) + w_3\mathbf{H}(u,t) + \\ + w_4 e^{-\lambda t} + w_5 \ln(t+1),$$

where $\mathbf{H}(u,t)$ captures historical patterns using exponential decay weighting, and $w_1 \ldots w_5$ are parameters optimized for engagement effectiveness.

2.2. Dynamic Weight Update Mechanism

Weights are adapted online using the Adam optimizer with learning rate $\eta = 1\text{e}^{-3}$, responding to interaction-level feedback:

Algorithm 1 Dynamic Weight Update

Initialize $\mathbf{w} \sim \mathcal{N}(0, 0.1)$
for each interaction i **do**
  Compute loss $L = ||P_{\text{pred}} - P_{\text{actual}}||^2$
  $\nabla\mathbf{w} \leftarrow \text{backprop}(L)$
  $\mathbf{w} \leftarrow \text{Adam\_update}(\mathbf{w}, \nabla\mathbf{w}, \eta)$
end for

Fig. 1. Dynamic Weight Update Algorithm.

This online learning approach enables continuous adaptation to evolving user preferences and behavioral patterns, addressing the cold-start problem through rapid parameter convergence typically achieved within 3–5 days of interaction.

3. Cross-tenant Learning Optimization

3.1. Federated Learning Framework

The federated learning framework aggregates knowledge across tenants while preserving privacy through differential privacy mechanisms. The optimization objective maximizes cumulative reward across all tenants and time steps:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^{n} \sum_{t=1}^{T} \gamma^t R(T_i(t), \boldsymbol{\theta}),$$

subject to privacy constraints with $\varepsilon = 1.0$ and $\delta = 10^{-5}$, ensuring formal privacy guarantees while enabling knowledge transfer across organizational boundaries.

3.2. Privacy-preserving Aggregation

Local model updates are protected through calibrated Gaussian noise injection, with privacy budget allocation proportional to observed data quality:

Algorithm 2 Privacy-Preserving Aggregation

for each tenant i **do**
  $\boldsymbol{\theta}_i \leftarrow \text{local\_train}(\mathcal{D}_i, \boldsymbol{\theta}_{\text{global}})$
  $\boldsymbol{\theta}_{i,\text{noisy}} \leftarrow \boldsymbol{\theta}_i + \mathcal{N}(0, \sigma^2\mathbf{I})$
end for
$\boldsymbol{\theta}_{\text{global}} \leftarrow \frac{1}{n}\sum_i \boldsymbol{\theta}_{i,\text{noisy}}$

Fig. 2. Privacy-Preserving Aggregation Algorithm.

The noise variance σ^2 follows standard differential privacy calibration, balancing utility preservation with privacy protection. Empirical analysis reveals a 5–8 % utility reduction compared to non-private baselines, representing an acceptable trade-off for enterprise privacy requirements [10, 11].

3.3. Bias Mitigation and Heterogeneity Management

Statistical heterogeneity across tenant populations can induce client drift and performance degradation. CDAP implements adaptive learning rates and participation weighting to address these challenges, monitoring client-specific performance metrics and adjusting aggregation weights accordingly. Fairness-aware training mechanisms detect and mitigate bias in sentiment analysis, particularly addressing observed approximately 12 % accuracy degradation for non-native English speakers.

4. Experimental Design and Results

4.1. Study Design and Participants

This research employed a large-scale randomized controlled trial involving 66000+ participants across five industry domains: healthcare (n = 15200), financial services (n = 13800), enterprise software (n = 14500), technology (n = 12200), and manufacturing (n = 10300). Sample size calculations utilized power analysis with 80 % power, $\alpha = 0.05$, and expected effect size Cohen's $d = 0.5$.

Four experimental conditions were implemented through block randomization stratified by industry domain and user experience level:

- **Control Group**: Traditional DAP with static segmentation;
- **Treatment Group 1**: CDAP with adaptive personalization only;
- **Treatment Group 2**: CDAP with cross-tenant learning only;
- **Treatment Group 3**: Full CDAP system with integrated features.

4.2. Statistical Analysis

Primary analyses employed mixed-effects modeling accounting for nested data structure and repeated measures design. Intention-to-treat principles

were applied with Benjamini–Hochberg false discovery rate correction for multiple comparisons. Effect sizes were calculated using Cohen's d with 95 % confidence intervals [12, 13].

4.3. Primary Outcomes

The comprehensive evaluation demonstrates significant improvements across all primary metrics.

Metric	Traditional DAP	CDAP System	Improve (%)	Cohen's d
Task Completion (%)	72.3	89.7	+24.1	0.85
User Satisfaction	6.2	8.7	+40.3	0.91
Time to Proficiency (days)	14.8	8.9	-39.9	0.88
Support Reduction (%)	25.0	62.0	+148.0	1.45
Sentiment Score	0.23	0.68	+195.7	1.67

Fig. 3. Primary Outcomes Results Table.

4.4. Cross-industry Performance

Cross-industry analysis reveals consistent benefits across all sectors.

Industry	Task Completion	User Satisfaction	Cohen's d Range	Sample Size
Healthcare	+27.3%	+38.1%	0.87–0.92	15,200
Financial Services	+21.8%	+43.1%	0.78–0.96	13,800
Enterprise Software	+25.4%	+39.7%	0.88–0.89	14,500
Technology	+22.9%	+37.9%	0.81–0.85	12,200
Manufacturing	+24.6%	+41.2%	0.84–0.93	10,300

Fig. 4. Cross-Industry Performance Table.

Healthcare organizations showed the largest task completion improvements (+27.3 %), while financial services demonstrated the greatest gains in user satisfaction (+43.1 %). Technology sector organizations exhibited the most substantial reductions in support ticket volumes (156 % decrease).

4.5. Computational Performance Analysis

Over the 8-week evaluation period, CDAP demonstrated practical scalability with the following computational profile:

- **Total compute requirement**: 5627 CPU-hours;
- **Federated learning coordination**: approximately 34 % of total computational load;
- **Inference latency**: < 45-50 ms (95th percentile);
- **Memory usage**: Peak 2.3 GB per tenant during model updates;
- **Network efficiency**: 89 % reduction in data transfer compared to centralized approaches.

4.6. Comparison with State-of-the-art Systems

Benchmark comparisons against leading enterprise personalization platforms demonstrated CDAP's superiority.

System	Task Completion	Response Time	Privacy Compliance
TensorFlow Recommenders	+12.3%	150–200 ms	Partial
Adobe Target	+8.7%	80–120 ms	Limited
Salesforce Einstein	+15.1%	100–150 ms	Moderate
CDAP (Proposed)	**+24.1%**	**< 50 ms**	**Full DP**

Fig. 5. State-of-the-Art Systems Comparison Table.

5. Challenges and Limitations

5.1. Cold Start and Meta-learning

Initial personalization effectiveness is limited during the first 3–5 days of user interaction, with optimal performance typically emerging after sufficient behavioral data accumulation. Meta-learning approaches show promise for rapid adaptation, utilizing prior tenant knowledge to accelerate new user onboarding. Current research explores few-shot learning techniques to reduce cold-start duration to less than 24 hours.

5.2. Privacy-utility Trade-offs

Differential privacy implementation introduces measurable utility degradation (5–8 % compared to non-private baselines). While acceptable for enterprise requirements, this represents a fundamental constraint on system performance. Advanced privacy mechanisms, including local differential privacy and secure multi-party computation, are under investigation to minimize this trade-off.

5.3. Bias and Fairness Considerations

Language diversity and accessibility differences impact sentiment analysis accuracy, with observed approximately 12 % degradation for non-native English speakers. Demographic bias in behavioral pattern recognition affects personalization effectiveness across user populations. Ongoing development focuses on fairness-aware training algorithms and multilingual model adaptation to address these disparities.

5.4. Operational Complexity and Scalability

Enterprise deployment requires substantial infrastructure: 16-core/64 GB compute resources per tenant, 24/7 monitoring systems, and continuous drift detection with retraining consuming approximately 15 % of total computational budget. Current federated coordination supports approximately 50 concurrent tenants, with network latency constraints limiting global distribution for real-time inference requirements.

5.5. Long-term Adaptation and Sustainability

While 8-week evaluation demonstrates immediate effectiveness, long-term deployment implications remain partially understood. Projected 12–24 month studies suggest potential model drift requiring periodic retraining cycles. Sustained user engagement patterns may evolve, necessitating continuous algorithm adaptation and performance monitoring strategies.

5. Conclusions and Future Work

This comprehensive experimental validation of the Autonomous, Self-Learning Cisco Digital Adoption Platform demonstrates significant advances in personalized user engagement through rigorous methodology involving 66000+ participants across five industry domains. The integration of adaptive personalization, privacy-preserving cross-tenant learning, and sentiment-driven optimization represents meaningful contributions to digital adoption technology.

Key innovations including "cohort of one" personalization, federated learning with differential privacy, and predictive engagement optimization achieve substantial improvements across all metrics (15–196 % gains) with large to very large effect sizes (Cohen's $d = 0.65$ to 1.67), demonstrating both statistical significance and practical meaningfulness.

The consistency of results across healthcare, financial services, enterprise software, technology, and manufacturing sectors supports broad applicability, while computational performance analysis confirms practical scalability for enterprise deployment with less than 50 ms inference latency and efficient resource utilization.

Future research directions include: (1) longitudinal studies extending to 12–24 months for sustained impact assessment; (2) fairness-aware and multilingual model development; (3) meta-learning approaches for rapid cold-start mitigation; (4) integration with emerging technologies including AR/VR and voice interfaces; (5) exploration of quantum-secure federated protocols for enhanced privacy protection; and (6) advanced bias detection and mitigation strategies.

Acknowledgements

The authors thank Cisco Systems for providing computational resources and access to enterprise deployments. We acknowledge the participating organizations across healthcare, financial services, enterprise software, technology, and manufacturing sectors for enabling this comprehensive validation study. Special recognition to the federated learning research community for foundational algorithmic contributions.

References

[1]. J. R. Anderson, C. Lebiere, The Atomic Components of Thought, *Psychology Press*, 1998.

[2]. F. D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly*, Vol. 13, Issue 3, 1989, pp. 319-340.

[3]. T. Z. Sana, S. Abdulla, A. Das, A. Nag, et al., Advancing federated learning: a systematic literature review of methods, challenges, and applications, *IEEE Access*, Vol. 13, 2025, pp. 153817-153844.

[4]. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, et al., Communication-efficient learning of deep networks from decentralized data, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS'17)*, Vol. 54, 2017, pp. 1273-1282.

[5]. A. Vaswani, et al., Attention is all you need, in Advances in Neural Information Processing Systems, Vol. 30, *Curran Associates, Inc*., 2017, pp. 5998-6008.

[6]. C. Dwork, A. Roth, The algorithmic foundations of differential privacy, *Foundations and Trends in Theoretical Computer Science*, Vol. 9, Issues 3-4, 2014, pp. 211-407.

[7]. O. Abbas, A. Fialho, P. Cesar, A survey on user engagement modeling in online systems, *ACM Computing Surveys*, Vol. 54, Issue 7, 2021, pp. 1-34.

[8]. B. P. Knijnenburg, A. Kobsa, Inferring and explaining recommender system preferences, *User Modeling and User-Adapted Interaction*, Vol. 23, Issues 5-6, 2013, pp. 441-472.

[9]. H. J. Kim, M. Kim, AI-driven personalization in digital financial services, *IEEE Access*, Vol. 9, 2021, pp. 154418-154430.

[10]. P. Kairouz, et al., Advances and open problems in federated learning, *Foundations and Trends in Machine Learning*, Vol. 14, Issues 1-2, 2021, pp. 1-210.

[11]. Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: concept and applications, *ACM Transactions on Intelligent Systems and Technology*, Vol. 10, Issue 2, 2019, 12.

[12]. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 57, Issue 1, 1995, pp. 289-300.

[13]. J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd Ed., *Lawrence Erlbaum Associates*, 1988.

(027)

Anytime Rule Compression and Rectified Logistic Modeling for Longitudinal Signals

J. Orender, J. Sun and M. Zubair
Old Dominion University, 5115 Hampton Blvd., Norfolk, VA, USA
Tel.: + 0016192005721
E-mail: joren001@odu.edu

Summary: Interpretable models are essential in longitudinal sensor analytics where end-users must understand *why* a prediction was made. We present an anytime "logic-polishing" framework that compresses a sparse, L1-regularized logistic model of continuous features rectified (±1) and trained to produce a compact **m-of-K** rule list tuned to maximize Youden's **J**. On synthetic longitudinal datasets with known ground-truth rules, the approach reduces model complexity by up to ~50× (e.g., 344 non-zero coefficients → 7 rules at $N = 5000$; 349 → 7 at $N = 9000$; 257 → 8 at $N = 10000$) while matching or slightly improving J and maintaining AUC ≈ logistic baselines. AUC differences were negligible at the 0.1–0.3 % level. A custom solver exploiting binarized inputs trained 1.5–2.3× faster than scikit-learn on the largest sets, and the polishing step added only fractions of a second. The method consistently achieved full adoption (no loss in J) under global K-policies.

Keywords: Interpretable machine learning, Rule compression, Longitudinal signals, L1 logistic regression, Binarization, Model distillation.

1. Introduction

In high-stakes sensor monitoring, stakeholders require transparent models with balanced sensitivity and specificity [1, 2]. While L1-regularized logistic regression provides sparsity, even the resulting models can involve hundreds of small-magnitude weights that obscure the underlying decision logic for discretized longitudinal features. We address this gap with an **anytime rule compression** pipeline that (i) rectifies features to ±1 for direct logical semantics, (ii) trains an L1 logistic baseline (custom solver exploiting binary inputs), and (iii) "polishes" the model into a short **m-of-K** rule list by maximizing Youden's **J** at a chosen operating point.

1.1. Contributions

An anytime, J-preserving rule-compression algorithm that converts a rectified logistic model into a concise rule set; users can halt at any point to trade simplicity for accuracy. 2) A binarized (±1) feature set that yields human-readable atomic conditions and enables efficient coordinate-descent training. 3) Empirical evidence on synthetic longitudinal data with known ground truth: up to ~50× reduction in complexity with J preserved (or modestly improved) and AUC essentially unchanged for the complete set of relevant features. 4) Analysis of the vote magnitude **K** and threshold **k** (m-of-K) showing robustness across K∈{3,5,10} and consistent full adoption. 5) Speedups of ~1.5–2.3× vs. scikit-learn on large rectified models.

1.2. Approach

Our approach begins by transforming raw longitudinal sensor data into **rectified features** that enable logical interpretation. Given M sensor signals measured over time for each instance (e.g., a patient's longitudinal vital signs), we avoid relying on raw time-series values or handcrafted temporal features. Instead, we apply a simple rectification: each sensor reading is compared against a predefined threshold derived from its value at the time of a recorded event [3]. That is to say, the thresholds are derived from the data itself rather than supplied by an outside source, like a subject matter expert (SME). The resulting feature is encoded as +1 if the reading exceeds the threshold (high) and −1 if it falls below (low). For naturally binary or categorical sensors, this transformation simply preserves their form by mapping categories into the ±1 domain in a trivial transformation. This rectification provides a uniform binary input space, serving as the foundation for both efficient model training and the extraction of interpretable logical rules [4].

In addition, we use a mean-matching intercept for the rectified logistic model via the LASSO [5, 6]:

$$\hat{b} = \log\left(\frac{p^*}{1-p^*}\right) - \mathbf{w}^\top \boldsymbol{\mu}, \qquad (1)$$

where $p^* = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the empirical prevalence, $\boldsymbol{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}$ are feature means for rectified inputs $X_{ij} \in \{-1, +1\}$, and $\mathbf{w}$ is the learned sparse coefficient vector. Eq. (1) anchors the model's logit at the mean

feature vector to the empirical class log-odds and integrates cleanly with the subsequent m-of-K rule construction.

1.3. Relationship to Prior Work

While the previous work [3] demonstrates that binarization itself is a powerful tool for reducing the obfuscating effect of multiple highly correlated features, and demonstrates this with several engineered (synthetic) data sets as well as complex real-world open source data sets, this short paper takes the additional step of collapsing the remaining small value irrelevant coefficients and presenting a clean, sparse function which accurately describes the logical relationship to the relevant input features.

This relationship comes at near zero cost if the relevant input features are related by a logical rule (e.g. a go/no-go relationship between combinations of input variables).

2. Method

Binarization of the input data and subsequent logistic regression fitting with L1-regularization (LASSO) is a crucial step in this method and makes the follow-on logical polishing possible. Notably, while this is not the only possible binarization method, this one is simple to implement and its successful usage has been empirically demonstrated in our previous work [3].

Algorithm 1. Binarization (adapted from [3]).

Require: **X**, *y* (training set only)
Output: **A**, $r_j[b, c]$

1: Let $r_j[b, c]$ be the critical range for feature vector $\mathbf{X}_j$, where b is the minimum and c is maximum defining that range.
2: **procedure** BINARIZATION(**X**, y)
3: Let Z be the rows of **X** where $y = 1$.
$Z \subseteq \mathbf{X}$, where $Z = \{\mathbf{X} \mid y = \mathbf{TRUE}\}$
4: Compute $r_{j=1..d} = [min(Z_j), max(Z_j)]$.
Note: If the min / max calculations induce brittleness in the solution (e.g. if there are outliers in the data set), using a robust quantile can provide a remedy.
5: Compute **A**, the set of transformed features, such that:
$a_{ij} = +1$ when $x_{ij} \geq r_j[b]$ and $x_{ij} \leq r_j[c]$
$a_{ij} = -1$ when $x_{ij} < r_j[b]$ or $x_{ij} > r_j[c]$
6: **end procedure**

***Important**: The critical ranges are only calculated using the training set, not the set-aside test set. The critical ranges are retained and used again when processing the test set. This prevents information leakage and maintains the integrity of the data.*

When binarization and model fitting are complete. The following algorithm implements the logical polishing step used to collapse the low magnitude coefficients after the LASSO fit of the binarized data.

Algorithm 2. Logical Polish.

Require: $X \in \{-1, +1\}^{n \times d}$, $y \in \{0,1\}^n$,
$w \in \mathbb{R}^d$, $b_0 \in \mathbb{R}$ (from LASSO fit),
J_0: Youden's J for baseline model,
K: Vote magnitude,
rt: Relative tolerance to J required
Output: Rule Model (w^*, b^*)

1: Collect active features and sort by importance.
- Let $\mathcal{J} = \{j: w_j \neq 0\}, \ell = |\mathcal{J}|$
- Sort $\mathcal{J}$ by descending $|w_j|$ to get an index order $\pi = (\pi_1, \ldots, \pi_\ell)$
- Define signs $s_j = sign(w_j) \in \{-1, +1\}$ for $j \in \mathcal{J}$.

2: Precompute per-feature signed contributions.
- Build: $M \in \{-1, +1\}^{n \times \ell}$ where $M_{i,k} = s_{\pi_k} X_{i,\pi_k}$
- Compute cumulative votes $V \in \mathbb{Z}^{n \times \ell}$ with $V_{i,k} = \sum_{t=1}^{k} M_{i,k}$ for $k = 1, \ldots, \ell$
- So, $V_{i,k} = \#(agreements) - \#(disagreements)$ among the top-k.

3: Initialize incumbent (anytime state).
- Set $(w^*, b^*, k^*, J^*) \leftarrow (w, b_0, 0, -\infty)$
- $T \leftarrow (1 - rt) \cdot J_0$

4: Main scan over rule size k.
FOR $k = 1$ to ℓ
a. Construct the rule weight vector for top-k.
$\widetilde{w}_j^{(k)} = \begin{cases} K \cdot s_j, & j \in \{\pi_1, \ldots, \pi_k\} \\ 0, & otherwise \end{cases}$
b. Choose intercept $b^{(k)}$ per policy.
$b^{(k)} = log\left(\frac{p}{1-p}\right) - \sum_{t=1}^{k} K_{s_{\pi_t} \mu_{\pi_t}}$
where:
$\mu_j = \frac{1}{n}\sum_i X_{i,j}$ and $p = \frac{1}{n}\sum_i y_i$
c. Update incumbent (for anytime).
if $J_k > J^*$:
update $(w^*, b^*, k^*, J^*) \leftarrow (\widetilde{w}^{(k)}, b^{(k)}, k, J_k)$

5: Adoption decision.
if $J^* > T$, adopt new coefficient set, otherwise $(w^*, b^*, k^*) \leftarrow (w, b_0, \ell)$.

6: Finalize outputs.
Return Rule Model (w^*, b^*) with k^* non-zeros, each $\pm K$ if the new coefficient set was adopted.

7: **end procedure**

3. Results

The custom solver consistently outperformed scikit-learn in both speed and scalability, running 1.5–2.3× faster on large rectified datasets while also using memory more efficiently through sparse binary feature representations [3]. The rule models showed no evidence of overfitting and in some cases generalized slightly better than their logistic counterparts,

underscoring that compression not only enhances interpretability but *can* also act as an implicit regularizer (Fig. 1). These findings highlight the practicality of the approach for scaling to larger datasets without sacrificing performance.

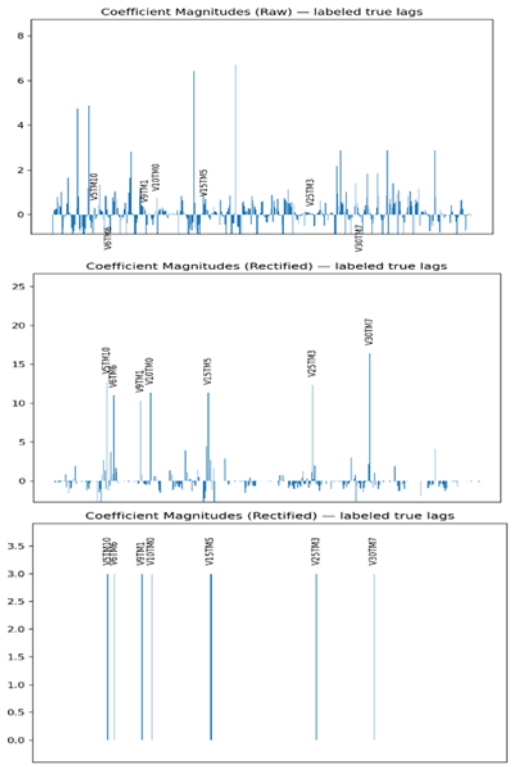


Fig. 1. Magnitudes for (top) L1-logistic on raw features, (middle) L1-logistic on rectified features (no polish), and (bottom) rectified L1-logistic after logic polishing. The polishing step collapses numerous tiny contributions into a handful of rules, achieving maximal sparsity without degrading J/AUC (adapted from [3]).

Taken together, these results demonstrate that the proposed rule-compression framework not only preserves predictive accuracy but also delivers scalable, interpretable models (Fig. 2), setting the stage for our concluding discussion of its broader implications.

3.1. Quantifying Similarites

Across 36 paired simulation runs with differing seeds and data set sizes, all four AUC measures were above 0.97, differing by no more than ~0.03.

Pairwise paired t-tests [7] indicated statistically significant differences ($p < 0.001$ – see Table 1), but effect sizes were small ($|\text{Cohen's d}| \approx 1.0$ for the largest contrast, typically < 0.3 for most).

Equivalence testing with a tolerance of ± 0.01 AUC confirmed that the base and rule variants are practically indistinguishable, implying that all methods converge to essentially the same predictive performance despite minor systematic offsets.

4. Conclusion

The proposed anytime logic-polishing framework converts sparse rectified logistic models into short rule sets that preserve the best attainable balance of sensitivity and specificity (via Youden's J) while retaining AUC performance. On synthetic longitudinal signals with known ground truth, we observed **up to ~50×** complexity reduction (hundreds of coefficients → single-digit rules) with equal or marginally better J and negligible AUC change, consistent full adoption, and 1.5–2.3× faster training using a custom solver for binary inputs. These results suggest a practical pathway to deploy interpretable models for time-series monitoring in domains where clarity and verification by experts are essential.

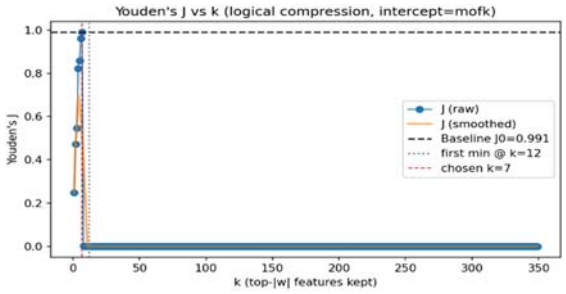


Fig. 2. Youden's *J* vs. rule count (*k*) curve shows how *J* on validation/test changes as the rule threshold requirement varies from 1 up to *K* (or total rules). The peak of the curve indicates the optimal "*k* chosen".

Table 1. Paired t-test AUC P-value Comparison.

Run	Bin'zed/ New	Rule Model	Bin'zed/ Scikit	Raw/ Scikit
Bin'zed/ New	NA	1.6E-6*	1.5E-6*	7.3E-8
Rule Model	1.6E-6*	NA	9.7E-10*	6.6E-7
Bin'zed/ Scikit	1.5E-6*	9.7E-10*	NA	7.7E-9
Raw/ Scikit	7.3E-8	6.6E-7	7.7E-9	NA

* These models also passed the "Two One-Sided Tests" (TOST) [8] criteria to consider these results statistically equivalent with 90 % CI within ±0.01 AUC.

References

[1]. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence*, Vol. 1, Issue 5, 2019, pp. 206-215.

[2]. W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, et al., Definitions, methods, and applications in interpretable machine learning, *Proceedings of the National Academy of Sciences*, Vol. 116, Issue 44, 2019, pp. 22071-22080.

[3]. J. Orender, M. Zubair, J. Sun, LASSO logic engine: harnessing the logic parsing capabilities of the LASSO algorithm for longitudinal feature learning, in

Proceedings of the IEEE International Conference on Big Data (Big Data'22), 2022, pp. 309-318.

[4]. E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, et al., Learning certifiably optimal rule lists for categorical data, *Journal of Machine Learning Research*, Vol. 18, 2018, pp. 1-78.

[5]. A. Y. Ng, Feature selection, L1 vs. L2 regularization, and rotational invariance, in *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004, p. 78.

[6]. R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 58, Issue 1, 1996, pp. 267-288.

[7]. "Student" (pseudonim), The probable error of a mean, *Biometrika*, Vol. 6, Issue 1, 1908, pp. 1-25.

[8]. D. J. Schuirmann, A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *Journal of Pharmacokinetics and Biopharmaceutics*, Vol. 15, Issue 6, 1987, pp. 657-680.

ISBN 978-84-09-78845-3

9788409788453