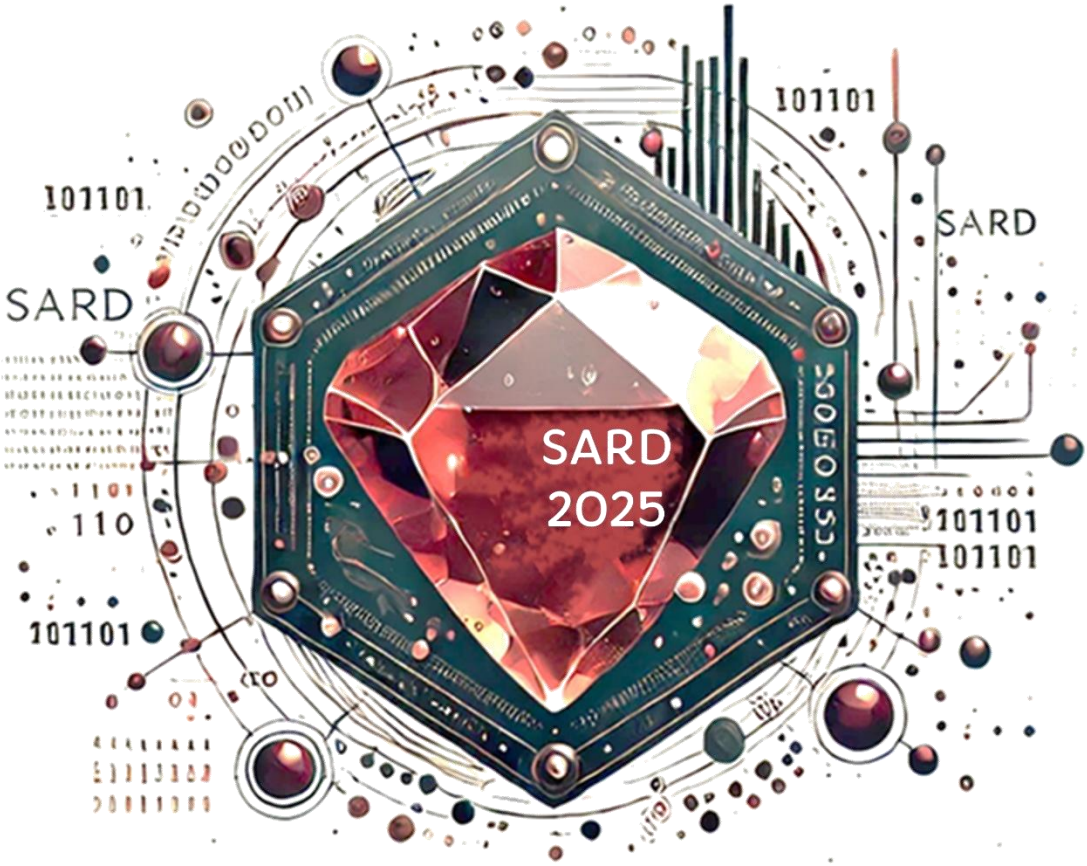
Proceedings of the
Seidenberg Annual Research Day

December 5 2025



SARD 2025
Seidenberg Annual Research Day

Hosted by
Seidenberg School of Computer Science and Information Systems, Pace University
New York, New York

edited by Li-Chiou Chen
Sung-Hyuk Cha

Proceedings of the Seidenberg Annual Research Day 2025

Editors:

Li-Chiou Chen	Pace University
Sung-Hyuk Cha	Pace University

Program Committee:

Miguel	Pace University
Soheyla Amirian	
Krishna Bathula	Pace University
Sara Falcone	Pace University
Yegin Genc	Pace University
Chienting Lin	Pace University
Miguel Mosteiro	Pace University
Juan Shan	Pace University
Namchul Shin	Pace University
Lixin Tao	Pace University

Hosted by:

The Seidenberg School of Computer Science and Information Systems

Published by Pace University Press

Available online
http://csis.pace.edu/∼scha/SARD2025

Table of Contents

Openning Ceremony

Invited Talks

Poster Competition

Welcome to the Seidenberg Annual Research Day 2025

Li-Chiou Chen
Dean of Seidenberg School of Computer Science & Information Systems

It is with great pleasure that we welcome you to the Seidenberg Annual Research Day (SARD) 2025, an esteemed conference hosted by the Seidenberg School of Computer Science and Information Systems at Pace University. This event has established itself as a cornerstone of academic excellence, celebrating both pioneering research and exceptional student contributions.

SARD 2025 offers a vibrant platform for faculty, students, and professionals to engage in meaningful dialogue, share groundbreaking ideas, and explore advancements in Computer Science, Data Science, and Information Technologies. By featuring a diverse range of participants, including seasoned researchers and emerging scholars, this conference fosters an inclusive environment that bridges academic and professional communities.

Expert talks by distinguished faculty members aim to inspire and challenge our attendees, ensuring accessibility and relevance for a broad audience, particularly students. This year's invited speakers are Dr. Sara Falcone and Dr. Miguel Mosteiro. We extend our sincere gratitude to them for their invaluable contributions.

The proceedings of SARD 2025 include 23 outstanding abstracts selected from over 26 submissions. These abstracts represent original research and thought-provoking problems pertinent to Computer Science and Information Systems. We commend all contributors for their dedication and innovation and thank them for enriching this conference with their work.

We hope this event serves as a catalyst for new ideas, collaborations, and learning opportunities. Thank you for your participation, and we look forward to your active engagement in the sessions ahead.

Welcome to SARD 2025!

A Multidisciplinary Investigation to Unravel the Complexity of the Sense of Embodiment in Teleoperation

Sara Falcone
Department of Computer Science, Pace University New York, NY
sfalcone@pace.edu

The sense of embodiment - the feeling that a virtual, robotic or remote body is one's own - is increasingly recognized as a critical factor in effective teleoperation. While robotics research has traditionally focused on improving control fidelity and stability, less attention has been devoted to understanding how perceptual, cognitive, and social dimensions jointly shape operators embodied experience. This talk presents a multidisciplinary investigation into the complexity of embodiment in teleoperation, integrating insights from robotics, cognitive science, and human"computer/robot interaction. Drawing on recent experimental findings, I will discuss how components such as ownership, agency, synchrony, and tactile feedback contribute to the formation of embodiment, and how these cues can be aligned with task goals to enhance performance and motor learning. By bridging technical approaches with theoretical frameworks, the work aims to identify principles that can guide the design of next-generation teleoperation systems; systems that not only extend human action at a distance but also foster a more intuitive and engaging embodied experience.

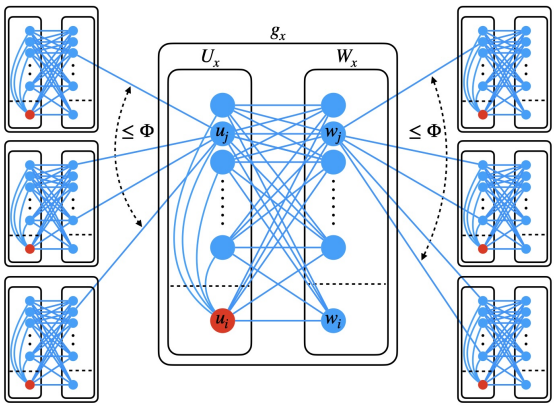
Deterministic Local Problems in Radio Networks: On the Impact of Local Domination and a Bit of Advice

Miguel A. Mosteiro
Pace University, Computer Science Department

The Radio Networks (RN) model was introduced about 40 years ago by Chlamtac and Kutten [1] as a multi-hop generalization of a multiple-access channel, attracting over the years a lot of attention. (A comprehensive list of work under this model would take many pages.) In the RN model of communication nodes can broadcast messages locally but their simultaneous transmissions can interfere with each other at their shared neighbors.

In this talk, I will overview our recent work [2] focusing on performing the very fundamental primitive of Local Broadcast, in spite of those interferences. In particular, I will discuss to what extent local knowledge, called *advice*, relating to the 2-local domination number γ_2 may speed up Local Broadcast. I will overview three $\widetilde{O}(\Delta\gamma_2^2)$-time Local Broadcast algorithms trading the level of adaptiveness (from oblivious to adaptive) for bits of advice per node (from $O(\log(\Delta\gamma_2))$ to 1).

I will also discuss how, using a complex adversarial network construction (see Figure), we can show that, for each quasi-adaptive deterministic Local Broadcast algorithm, there exists some RN that requires $\Omega(\min\{(\min\{\Delta,\gamma_2\}/\log n)^2, n\})$ communication rounds. This lower bound is *i*) nearly quadratically better than previous work for this class of algorithms, *ii*) it applies to a wider class of algorithms than previous work for fully oblivious, *iii*) it is similar than previous bound for a much less demanding problem, and *iv*) it takes into account the local domination parameter γ_2.



References

[1] I. Chlamtac and S. Kutten. On broadcasting in radio networks-problem analysis and protocol design. *IEEE Transactions on Communications*, 33(12):1240–1246, 1985.

[2] P. Garncarek, T. Jurdzinski, D. R. Kowalski, S. Kutten, and M. A. Mosteiro. Deterministic local problems in radio networks. In *Proc. of the 36th Int. Symp. on Algorithms and Computation (ISAAC)*, 2025. To appear.

Automated vs. Manual BML Volume Feature Assessment in Knee Replacement Surgery Prediction

Hetali Chavda[1], Khaing Su[1], Kevin Wang[2], Khushi Vithalani[1], Luca Pistella[1], Zender Aurelien-Nikolai[1], and Juan Shan[1]
[1]Department of Computer Science, Pace University New York, NY
[2]Hopewell Valley Central High School, Pennington, NJ
{hchavda, ks78979n, kv47141n, lp98547n, za51459n, jshan}@pace.edu,
{kevinwang}@hvrsd.org

Bone marrow lesions (BMLs) are important imaging biomarkers of knee osteoarthritis (OA), a progressive joint disease that often leads to pain, disability, and eventually knee replacement (KR) surgery. However, manual BML labeling is time-consuming and prone to inter-observer variability. In this study, we introduce an automatic MRI-based segmentation approach to extract BML volumes and assess whether these automated imaging features, together with clinical data, can improve prediction of KR surgery within five years. Two datasets were used: Dataset A (300 knees) for training and evaluating a BML segmentation model, and Dataset B (1,393 knees) for developing the KR prediction model using the features extracted by the trained segmentation network. For comparison, we also built a prediction model using manually labeled BML volume features to assess how well automated features perform relative to manual ones. Manual BML segmentations served as the ground truth for training and validating the segmentation models. Logistic regression was used for KR prediction, and clinical five-year follow-up data were used to determine whether each subject ultimately underwent KR surgery. Results show that although segmentation accuracy on Dataset A was moderate, the KR prediction model achieved substantially higher performance after replacing manual BML features with automated BML volumes, with AUC improving from 0.87 to 0.94. These findings indicate that automatically extracted imaging biomarkers, even when the segmentation is not perfect, can improve KR prediction and hold strong potential for large-scale OA progression studies.

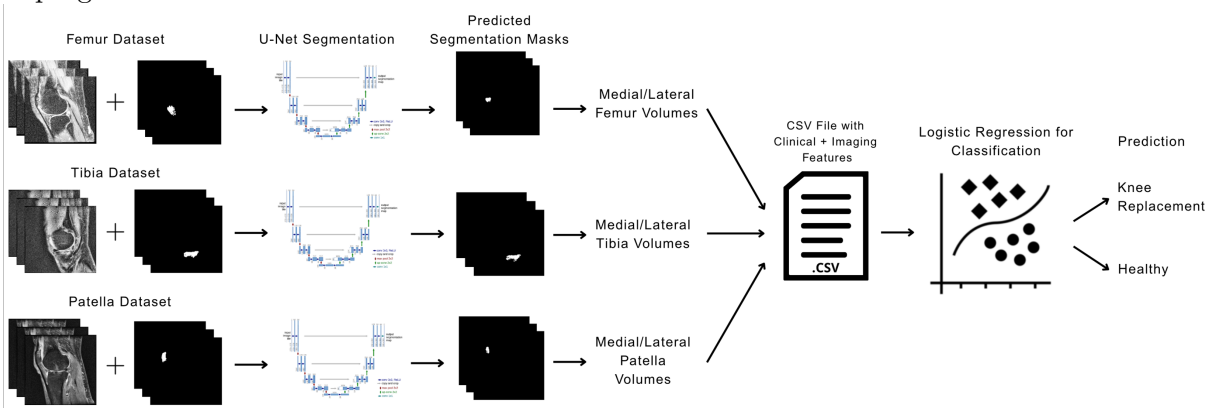


Figure 1: Overall Workflow

References

[1] R. Ponnusamy, M. Zhang, Y. Wang, X. Sun, M. Chowdhury, J. B. Driban, T. McAlindon, and J. Shan, *Automatic segmentation of bone marrow lesions on MRI using a deep learning method*, Bioengineering 2024, vol. 11, no. 4: 374.

Motion-Gated Vision Transformers for Efficient Real-Time Video Inference

Ruturaj Dixit and D. Paul Benjamin
Seidenberg School of Computer Science and Information Systems,
Pace University, New York, NY, USA
{rd85458n,dbenjamin}@pace.edu

Deployed vision transformers (ViTs) and vision–language models (VLMs) face a key bottleneck when operating on continuous video streams: every frame is treated as an independent image, leading to redundant computation and high latency. In this work we investigate a simple but effective *motion-gated inference* strategy for real-time video feeds. A lightweight motion proxy, based on frame differencing in a downsampled grayscale space, decides whether a new frame is sufficiently different from the last processed frame to warrant a transformer forward pass. When motion is small, the previous prediction is reused and the heavy backbone is skipped; when motion exceeds a tunable threshold, a ViT or OWL-ViT model is invoked normally.

We prototype this approach on a laptop GPU using live webcam input and evaluate five mainstream image transformers (ViT-Tiny/Small/Base and DeiT-Small/Base) as well as two open-vocabulary object detectors (OWL-ViT Base/Large). Across 400 real frames, our motion-gated scheduler skips 59–98% of frames, reducing *effective* per-frame latency by up to 97% while maintaining stable predictions on background-dominant scenes. For example, OWL-ViT Large improves from $\sim$ 1 FPS to $\sim$ 2.6 FPS, and image-classification ViTs achieve effective FPS in the 4k–5k range when most frames are static. We report detailed trade-offs between motion threshold, frame-skipping rate, and prediction stability, and outline how this idea can be combined with early exiting, token pruning, or hardware-aware kernels for further gains. The proposed motion-gated pipeline is model-agnostic, easy to integrate into existing video analytics stacks, and provides a practical baseline for future research on streaming-efficient transformers.

References

[1] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[2] H. Touvron *et al.*, "Training Data-Efficient Image Transformers & Distillation through Attention," in *Proc. International Conference on Machine Learning (ICML)*, 2021.

[3] M. Minderer *et al.*, "Simple Open-Vocabulary Object Detection with Vision Transformers," in *Proc. European Conference on Computer Vision (ECCV)*, 2022.

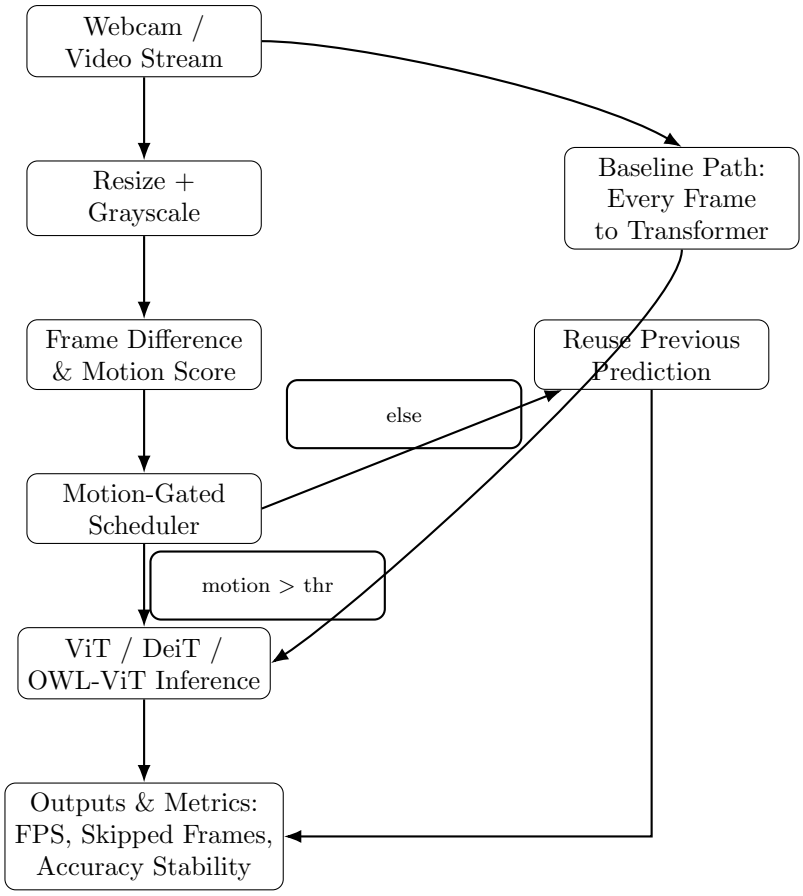


Figure 1: Motion-gated streaming pipeline. The baseline path sends every frame to the transformer, while the proposed path uses a lightweight motion proxy to skip redundant frames and reuse the last prediction when the scene is static.

Bootstrapping Multi-Modal Semantic-Vector Spaces

Watson Blair, J. Xavier, and Soheyla Amirian
Seidenberg School, Pace University, New York, NY, USA
{wb43827n, samirian}@pace.edu

We present an alignment and evaluation architecture for bootstrapping multimodal systems into a shared Semantic-Vector Space (SVS) (Fig. 1) that decouples encoding, inference, and decoding tasks so each module can be independently aligned and evaluated; our training pipeline leverages momentum-based cross-modal distillation as described in [1] to stabilize joint embedding learning across noisy cross-modal pairings. To minimize compute and data costs, we align pretrained encoders and decoders using lightweight adapter MLPs[2](Fig. 2) and synthesize cross-modal training pairs via backtranslation and contrastive objectives, which was shown by [3] to be effective for generating high-fidelity synthetic pairings and organizing cross-modal latent structure. We demonstrate a methodology for rapidly bootstrapping multimodal encoder and decoder swarms using open-weight vision, text, and audio models. This space can be used to further explore the utility of the Large Concept Model[4] and continuous-latent reasoning[5] approaches as well as exploring the applicability of statistical analysis and pattern recognition algorithms for exploring relationships in the semantic space. Our results provide a reproducible methodology for aligning new modalities to a Shared SVS, highlights practical trade-offs between adapter-based alignment and full-model fine-tuning, and show that a shared semantic-vector modality can serve as an efficient, extensible foundation for robust multimodal AI.

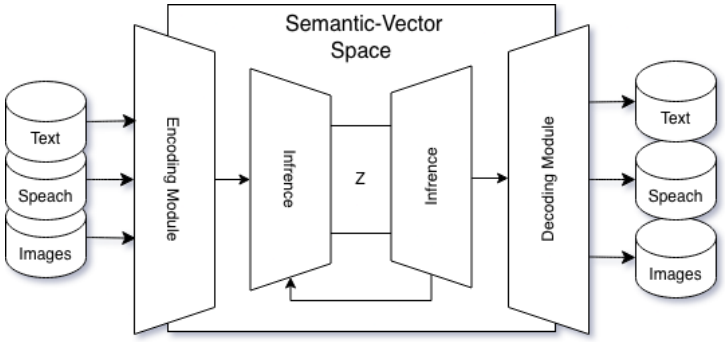


Figure 1: Semantic-Vector Space

References

[1] B. Zheng, M. Ablimit and A. Hamdulla, *"Cross-Modal Alignment for End-to-End Spoken Language Understanding Based on Momentum Contrastive Learning,'* in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Seoul, Korea, Republic of, 2024, pp. 10576-10580

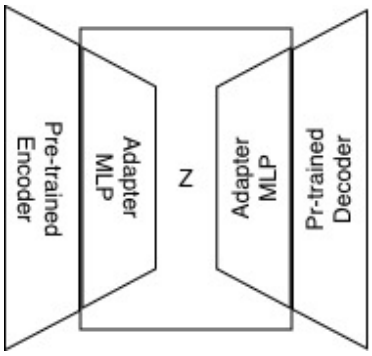


Figure 2: Adapter Architecture

[2] *"APE: Aligning Pretrained Encoders to Quickly Learn Aligned Multimodal Representations,,"* Apple Machine Learning Research, 2022. `https://machinelearning.apple.com/research/ape` accessed Nov. 23, 2025

[3] Lugosch, Loren, et al., *"Using Speech Synthesis to Train End-To-End Spoken Language Understanding Models,"* in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), May 2020, pp. 8499–8503

[4] L. Barrault, et al., *"Large Concept Models: Language Modeling in a Sentence Representation Space,"* FAIR at Meta, Dec. 12, 2024 `https://github.com/facebookresearch/large_concept_model`

[5] Hao, S, et al., *"Training Large Language Models to Reason in a Continuous Latent Space."* Meta AI, 3 Nov. 2025, `arxiv.org/abs/2412.06769` accessed 18 Nov. 2025.

A Multimodal Motion Dataset and Computational Analysis of Guangdong Lion Dance

Xianrong Liang[1], Lixin Tao[1], and Zhenghao Lu[2],
1 Computer Science Department, Pace University, New York, NY, USA
2 Guangdong Lion Dance Association in Tianhe District, Guangzhou, GD, China
{xliang, ltao}@pace.edu, {330270210}@qq.com

This paper presents the first multimodal motion dataset for Guangdong Lion Dance (广东醒狮), an important form of Chinese intangible cultural heritage characterized by expressive movements and percussion-driven rhythm. Using video-based pose extraction with BlazePose-full, we obtain 33 human landmarks and introduce Lion-11, a domain-specific skeletal abstraction capturing key lion components such as horn, nose, chin, shoulder, tail, and four paws. The dataset includes 3D trajectories, velocity–acceleration curves, and synchronized audio amplitude from drum, gong, cymbals, and wooden clappers. We further annotate and analyze the canonical Eighteen Styles, revealing movement–rhythm coupling patterns and distinct biomechanical signatures. Visualizations—such as 3D skeleton animations and trajectory heatmaps—demonstrate the interpretability of the representation. This dataset provides a foundation for research in digital heritage preservation, VR training systems, movement generation, and robotics imitation learning, bridging traditional cultural performance with modern computational methods.

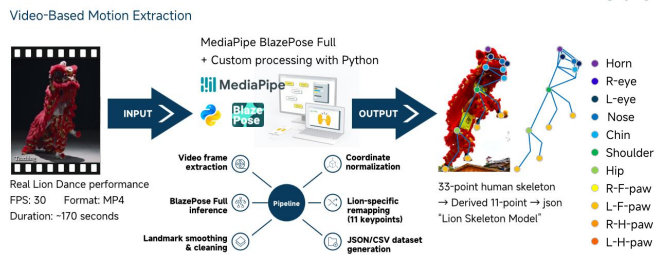


Figure 1: Video-Based Motion Extraction Pipeline.

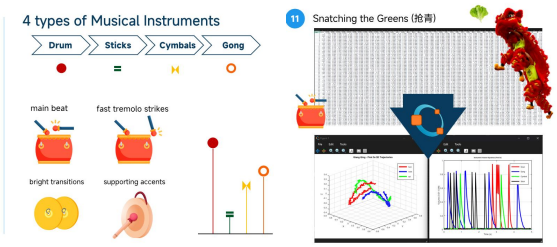


Figure 2: Motion and Audio Visualization.

References

[1] He, X. (2018). *Cantonese Lion Dance Culture and Its Historical Development*. Guangdong People's Publishing House.

[2] Dai, Y., & Zhang, L. (2020). "Cultural Symbolism and Performance Structure of Southern Lion Dance." Journal of Chinese Ethnic Arts, 12(3), 44–58.

[3] Bazarevsky, V. et al. (2020). "BlazePose: On-device Real-time Body Pose Tracking." Google Research.

[4] Zhou, F., & De la Torre, F. (2009). "Canonical Time Warping for Alignment of Human Behavior." NIPS, 2286–2294.

[5] Henter, G. E. et al. (2020). "MoDi: Motion Diffusion Model for Human Motion Generation." CVPR.

[6] Chen, M., Radford, A., et al. (2021). "Multimodal Sequence Learning for Motion and Audio." OpenAI Technical Report.

[7] Kenderdine, S. (2016). "Embodied Museography: Digital Performance and Cultural Heritage." Digital Scholarship in the Humanities, 31(2), 317–332.

An Advanced AI-Driven Database System

M. Tedeschi[1,3], S. Rizwan[1], C. Shringi[2], V. Devram Chandgir[2], and S. Belich[3],
[1]Seidenberg School, Pace University, New York, NY, USA
[2]Computer Engineering, NYU, New York, NY, USA
[3]Department of Computer Systems Technology, CityTech, CUNY, Brooklyn, NY, USA
{mtedeschi,sl00399n}@pace.edu,{cs7810,vc2651}@nyu.com,
{mtedeschi,serge.belich85}@citytech.cuny.edu

Contemporary database systems, while effective, suffer severe issues related to complexity and usability, especially among individuals who lack technical expertise but are unfamiliar with query languages like Structured Query Language (SQL). This paper presents a new database system supported by Artificial Intelligence (AI), which is intended to improve the management of data using natural language processing (NLP) - based intuitive interfaces, and automatic creation of structured queries and semi-structured data formats like yet another markup language (YAML), java script object notation (JSON), and application program interface (API) documentation. The system is intended to strengthen the potential of databases through the integration of Large Language Models (LLMs) and advanced machine learning algorithms. The integration is purposed to allow the automation of fundamental tasks such as data modeling, schema creation, query comprehension, and performance optimization. We present in this paper a system that aims to alleviate the main problems with current database technologies. It is meant to reduce the need for technical skills, manual tuning for better performance, and the potential for human error. The AI database employs generative schema inference and format selection to build its schema models and execution formats. This enables it to enhance performance continuously, for example, generative pretrained transformer 4 (GPT-4), in working with diverse types of databases such as relational, not only SQL or NoSQL, graph databases, and vector stores. In addition, reinforcement learning mechanisms are investigated to facilitate ongoing improvement and adaptation of performance. This research aims to render the technical expertise of the user unnecessary in performing elementary database operations. The article proposes overcoming the existing barriers by critical integration of advanced AI methods, i.e., LLMs and reinforcement learning, to create an auto-adapting, user-friendly database system. This integrated, AI-driven approach is new since it combines previously separate breakthroughs, employing modern machine learning methods to solve long-standing usability and performance issues in a unified manner. The proposed research follows different approaches to integrating state-of-the-art machine learning techniques, namely generative AI models, reinforcement learning for continuous system optimization, and relative performance comparisons. Analytical instruments include empirical case studies and comparative performance comparison with existing database technologies (SQL, NoSQL, NewSQL, graph databases) under various data scenarios. Additionally, approaches to detecting and mitigating AIspecific problems such as query hallucinations, schema drift, and ongoing schema evolution are explored and experimentally tested. Compared to existing solutions such as Massachusetts Institute of Technology (MIT's) GenSQL, which focuses on applying probabilistic models to statistical inference and analysis of table data, our solution is aimed at complete database management automation, with dynamic schema creation, adaptive performance optimization across heterogeneous data types and database systems, and natural language query processing. Further, our model explicitly integrates reinforcement learning mechanisms for continued self-improvement, addressing real-time schema evolution and query hallucinations issues that GenSQL does not explicitly handle.

References

[1] L. G. Azevedo, R. F. S. Souza, E. F. Soares, R. M. Thiago, J. C. C. Tesolin, A. C. Oliveira, and M. F. Moreno, *"A polystore architecture using knowledge graphs to support queries on heterogeneous data stores,"* in *Proceedings of the 34th International Conference on Database and Expert Systems Applications (DEXA)*, 2023. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-031-43153-1_27`. Accessed 7 May 2025.

[2] J. Fan, Z. Gu, S. Zhang, Y. Zhang, Z. Chen, L. Cao, G. Li, S. Madden, X. Du, and N. Tang, *"Combining small language models and large language models for zero-shot NL2SQL,"* *Proc. VLDB Endowment*, vol. 17, no. 11, pp. 2750–2763, 2024. Retrieved from `https://www.vldb.org/pvldb/vol17/p2750-fan.pdf`. Accessed 7 May 2025.

[3] J. Lu and I. Holubová, *"Multi-model databases: A new journey to handle the variety of data,"* *ACM Computing Surveys*, vol. 52, no. 3, Article 55, 2019. Retrieved from `https://doi.org/10.1145/3323214`. Accessed 7 May 2025.

[4] Z. H. Liu, J. Lu, D. Gawlick, H. Helskyaho, G. Pogossiants, and Z. Wu, *"Multi-model database management systems – a look forward,"* in *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, LNCS 11137, pp. 16–29, Springer, 2019. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-030-17488-0_2` and `https://www.cs.helsinki.fi/u/jilu/documents/Poly.pdf`. Accessed 7 May 2025.

[5] R. Marcus, P. Negi, H. Mao, N. Tatbul, M. Alizadeh, and T. Kraska, *"Bao: Making learned query optimization practical,"* in *Proc. ACM SIGMOD Int. Conf. Management of Data*, pp. 1275–1288, 2021. Retrieved from `https://doi.org/10.1145/3448016.3452838`. Accessed 7 May 2025.

[6] M. Ramadan, A. El-Kilany, H. M. O. Mokhtar, and I. Sobh, *"RL_QOptimizer: A reinforcement learning based query optimizer,"* *IEEE Access*, vol. 10, pp. 70502–70515, 2022. Retrieved from `https://doi.org/10.1109/ACCESS.2022.3187102`. Accessed 7 May 2025.

[7] Z. Yan, V. Uotila, and J. Lu, *"Join order selection with deep reinforcement learning: Fundamentals, techniques, and challenges,"* *Proc. VLDB Endowment*, vol. 16, no. 12, pp. 3882–3885, 2023. Retrieved from `https://www.vldb.org/pvldb/vol16/p3882-yan.pdf`. Accessed 7 May 2025.

[8] F. Ye, Y. Li, X. Wang, N. Nedjah, P. Zhang, and H. Shi, *"Parameters tuning of multi-model database based on deep reinforcement learning,"* *Journal of Intelligent Information Systems*, vol. 61, pp. 167–190, 2023. Retrieved from `https://doi.org/10.1007/s10844-022-00762-0`. Accessed 7 May 2025.

[9] F. Zhao, D. Agrawal, and A. El Abbadi, *"Hybrid querying over relational databases and large language models,"* in *Conf. Innovative Data Systems Research (CIDR)*, 2025. Retrieved from `https://www.cidrdb.org/cidr2025/papers/p57-zhao.pdf` and `https://vldb.org/cidrdb/papers/2025/p10-zhao.pdf`. Accessed 7 May 2025.

[10] X. Zhou, Z. Sun, and G. Li, *"DB-GPT: Large Language Model Meets Database,"* *Data Science and Engineering*, vol. 9, pp. 102–111, 2024. Retrieved from `https://doi.org/10.1007/s41019-023-00235-6`. Accessed 7 May 2025.

Building Data Pipelines Using Python: A Hands-On Journey in Web Scraping, Analysis, and Visualization

M. Tedeschi[1,2], M.G. Gattani[1], S. Sanjaybhai Nathani[1], H. Patel[1], S. Belich[2], and S. Gabani[1]
[1]Seidenberg School, Pace University, New York, NY
[2]College of Systems Technology, CityTech, CUNY, Brooklyn, NY
{mtedeschi, mg86111n, sn55541n, hp21565n, sg77891n}@pace.email.edu
{mtedeschi, serge.belich85}@citytech.cuny.edu

In today's data-centric world, the ability to extract, clean, analyze, and visualize real-world data is foundational in data analytics. As part of a hybrid-format academic course—combining alternating weeks of in-person and online instruction—we completed two major hands-on projects that developed practical skills in web scraping, data preprocessing, and visualization using Python. The hybrid learning environment encouraged technical experimentation, peer collaboration, and iterative learning. The first project, Basketball Data Scraping and Visualization, focused on sourcing historical NBA game statistics (2020–2024) from Basketball-Reference.com. Using Python libraries such as requests, Beautiful Soup, pandas, and matplotlib, we extracted structured HTML tables, cleaned the data into usable DataFrames, and created over 25 visualizations. These included bar charts, scatter plots, and line graphs that illustrated trends in team performance, scoring averages, seasonal fluctuations, and win-loss ratios across franchises. We refined the visual output through advanced styling, layout adjustments, and labeling techniques to ensure clarity and readability. The second project, Hacker News Web Scraper and Community Insights, addressed a platform with an inconsistent HTML structure and minimal markup. We built a scalable multi-page scraper capable of extracting post titles, ranks, vote counts, and comment counts. This required handling common scraping challenges, including HTTP 403 errors, bot detection, and irregular DOM elements, which we mitigated through adaptive scraping logic, dynamic headers, and error handling routines. Visualization of the extracted data was achieved using both seaborn and matplotlib, allowing us to present key engagement metrics through histograms, box plots, pie charts, and time-series graphs. Seaborn's high-level syntax facilitated rapid visualization, while Matplotlib offered granular control for polishing presentation elements. Both projects followed a unified pipeline: data acquisition, transformation, analysis, and presentation. We gained experience not only in technical scripting but also in designing scalable workflows and interpreting complex datasets. The contrast between the two projects: structured basketball data versus dynamic social platforms challenged us to adapt tools and strategies while preserving analytical rigor. The hybrid course structure further enhanced our learning by pairing independent development in online settings with collaborative troubleshooting during live sessions. This dual approach simulated real-world, distributed team environments, reinforcing self-directed learning and group problem-solving. Ultimately, these projects helped solidify our proficiency with essential data science tools and practices. Working extensively with Python libraries such as Beautiful Soup, pandas, requests, matplotlib, and seaborn deepened our understanding of the end-to-end data pipeline. More importantly, the experience strengthened our analytical thinking, error-handling strategies, and visual storytelling skills. These outcomes form a strong foundation for future work in areas such as real-time scraping, dashboard development, and applied machine learning

References

[1] L. Richardson, *Beautiful Soup 4.13.0 Documentation*. Crummy, 2025. Accessed September 18, 2025. Retrieved from `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`.

[2] J. D. Hunter, *Matplotlib: A 2D Graphics Environment*. IEEE Computer Society, 2007. Available from `https://matplotlib.org/stable/users/index.html`.

[3] M. Waskom, *Seaborn Documentation*. 2025. Accessed September 18, 2025. Retrieved from `https://seaborn.pydata.org/`.

[4] The pandas development team, *pandas 2.1.1 Documentation*. 2025. Accessed September 18, 2025. Retrieved from `https://pandas.pydata.org/docs/`.

[5] K. Reitz, *Requests Documentation*. 2025. Accessed September 18, 2025. Retrieved from `https://requests.readthedocs.io/en/latest/`.

[6] XlsxWriter Contributors, *XlsxWriter Documentation*. 2025. Accessed September 18, 2025. Retrieved from `https://xlsxwriter.readthedocs.io/`.

[7] Y Combinator, *Hacker News Website*. 2025. Accessed September 18, 2025. Retrieved from `https://news.ycombinator.com`.

[8] Basketball-Reference, *2020–21 NBA Season Summary*. 2025. Available from `https://www.basketball-reference.com/leagues/NBA_2021.html`.

[9] M. Tedeschi, S. Belich, P. Ricaurte Quijano, C. Danoff, J. Corneli, and S. Ayloo, *Peeragogy and Teaching*. In: *ICERI2024 Proceedings*. 2024. pp. 6079–6089.

[10] *Beautiful Soup Documentation*. Accessed September 20, 2024. Retrieved from `https://beautifulsoup-4.readthedocs.io/en/latest/`.

[11] *Python 3.13.7 Documentation*. Accessed September 18, 2025. Retrieved from `https://docs.python.org/3/`.

[12] *pandas Documentation*. Accessed September 20, 2024. Retrieved from `https://pandas.pydata.org/docs/index.html`.

[13] *Matplotlib 3.10.6 Documentation*. Accessed September 18, 2025. Retrieved from `https://matplotlib.org/stable/index.html`.

[14] *Allrecipes*. Accessed September 18, 2025. Retrieved from `https://www.allrecipes.com/`.

[15] M. Tedeschi, *From Classroom to Online Education – An Educator's Insights*. In: Proceedings of the 27th European Conference on Pattern Languages of Programs (EuroPLoP '22); 2022 Jul; Kloster Irsee, Germany. New York (NY): Association for Computing Machinery; 2023. Article 15, pp. 1–15. Available from `https://doi.org/10.1145/3551902.3551977`.

Guardrails for AI Agents: Detecting Jailbreaks, Misuse, and Cyberattack Automation in Large Language Models

F. Happymore Masoka, S. Chipo Makoni, T. Mitchel Kalowa, and Krishna Bathula
Seidenberg School of Computer Science and Information Systems
Pace University, NY, USA
{hm78761n, cm88450, mk91526, kbathula}@pace.edu

Advances in large language models (LLMs) have created new opportunities for attackers to automate complex cyber intrusions with minimal effort[1]. This project introduces a Cyber-AI Misuse Detection System designed to identify and flag malicious LLM activity in real time. The system employs a transformer-based classifier trained on curated datasets of benign, adversarial, and jailbreak prompts, enabling it to distinguish harmful intent with high precision[2]. A FastAPI-driven detection engine performs prompt preprocessing, embedding extraction, classification, anomaly scoring, and forensic logging. High-risk interactions are stored in MongoDB to support analysis of exploit attempts, evasion behaviors, and multi-step attack chains. A real-time dashboard visualizes threat patterns and system alerts to aid security monitoring. This work provides a scalable defensive framework for addressing emerging AI-enabled cyber threats and contributes to future AI governance, auditing, and safety infrastructure.

References

[1] L. Whittaker, "Anthropic warns of AI-driven hacking campaign linked to China," *Scripps News*, Feb. 2025. [Online]. Available: `https://www.scrippsnews.com/science-and-tech/artificial-intelligence/anthropic-warns-of-ai-driven-hacking-campaign-linked-to-china`

[2] M. Jones, "Artificial intelligence (AI)—The need for new safety standards and methodologies," *Journal of System Safety*, vol. 55, no. 3, pp. 11–22, 2020.

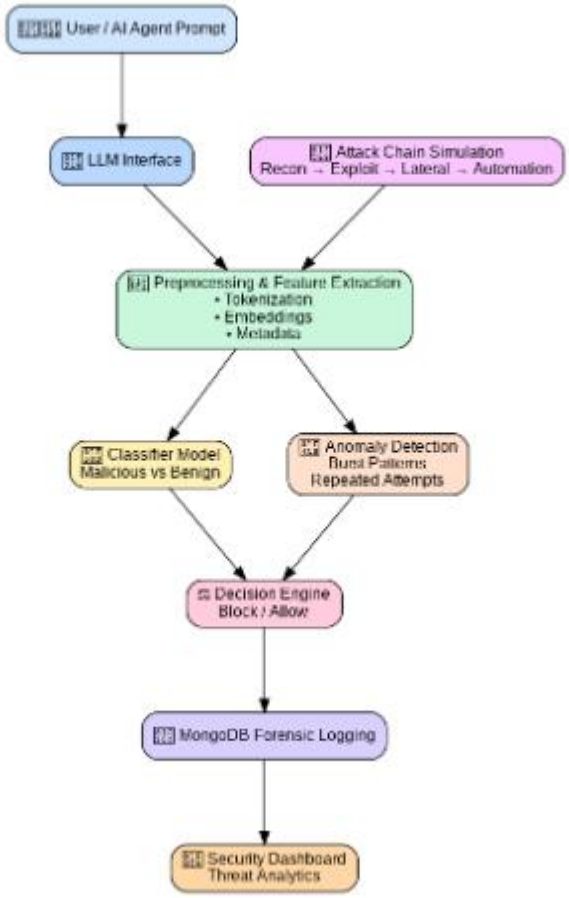


Figure 1: System Architecture Diagram for the Cyber-AI Misuse Detection Pipeline

FAIR-Agent: A Multi-Agent Framework for Quantifiably Trustworthy Large Language Models

Somesh Ghaturle, Darshil Malviya, Priyank Mistry, and Krishna Bathula
Seidenberg School of Computer Science and Information Systems
Pace University, NY, USA
{sg12345n, dm67890p, pm34567q, kbhatula}@pace.edu

Large Language Models (LLMs) such as ChatGPT, Claude, and Gemini demonstrate remarkable capabilities but lack quantifiable trustworthiness metrics, preventing adoption in regulated industries. We introduce FAIR-AGENT, the first LLM system with measurable trustworthiness through novel FAIR metrics: Faithfulness, Adaptability, Interpretability, and Risk-awareness. Our multi-agent architecture achieves a composite FAIR score of 62.0%, representing a 205% improvement over market-leading alternatives (average 25.0%). Through evidence-based responses from 53 curated sources including specialized datasets (FinQA, TAT-QA, MIMIC-IV, PubMedQA), chain-of-thought reasoning, and automated safety compliance, FAIR-AGENT delivers 100% source citation coverage versus 0-5% for competitors. Our dynamic baseline calculation system provides scientifically validated performance measurements, establishing the first quantifiable framework for trustworthy AI in healthcare and financial domains.

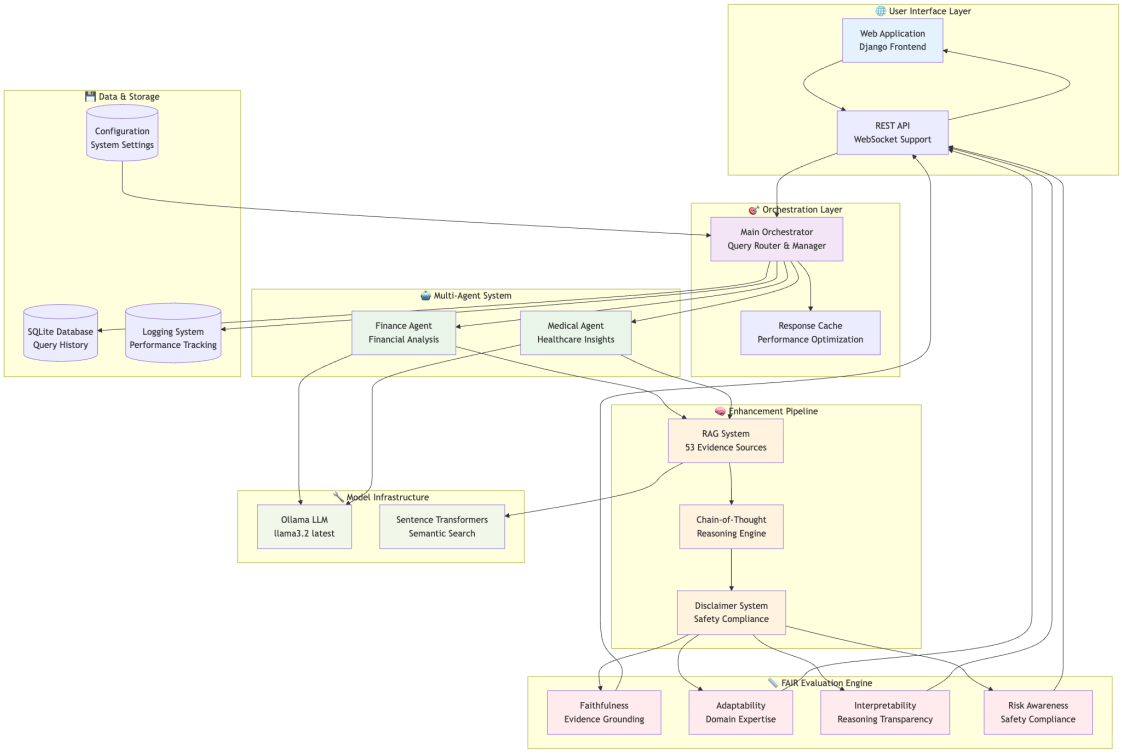


Figure 1: FAIR-Agent system architecture showing multi-agent coordination, evidence-based enhancement pipeline, and real-time FAIR evaluation engine.

BiasLens: An Explainability-Driven Framework for Bias Detection in Large Language Models

Isha Mukherjee, Kristina E Kim, and Krishna Bathula
Seidenberg School of CSIS, Pace University, New York City, NY, USA
im81479n@pace.edu, kk55355n@pace.edu, kbathula@pace.edu

This study presents an explainability-based framework for analyzing social bias in large language models (LLMs). We conduct a systematic comparison between two popular LLMs: GPT-4 Turbo and LLaMA 3 70B. Using two benchmark datasets, BOLD and StereoSet, our pipeline integrates bias evaluation and model explainability into a reproducible workflow. First, open-ended prompts and contexts of various domains from the datasets are used to elicit text generations from both models under deterministic settings. Next, we apply TextBlob and Detoxify to perform sentiment analysis and measure toxicity levels. Then, we introduce explainability through SHAP to identify token-level attribution. Additional metrics such as stereotype alignment, WEAT-style association tests, and trustworthiness metrics are used to quantify both bias magnitude and explanation reliability. Both models show subtle biases in polarity/subjectivity, generally low toxicity scores, and potentially biased language. Based on these insights, we propose a bias mitigation framework that uses these token-level explanations to filter or rephrase outputs and guide model fine-tuning with less biased data. By combining these token-level and global bias scores with trustworthiness metrics, this approach enables the systematic detection, explanation, and iterative reduction of bias, providing actionable insights for a safer and more equitable deployment of LLMs.

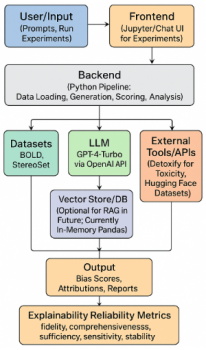


Figure 1: System architecture

References

[1] I. Gallegos, "Bias and fairness in large language models: A survey," *Comput. Linguist.*, vol. 50, no. 3, pp. 1097–1195, Sep. 2024, doi: 10.1162/coli_a_00524.

[2] Hu, T., Kyrychenko, Y., Rathje, S. *et al.* Generative language models exhibit social identity biases. *Nat Comput Sci* 5, 65–75 (2025), doi: 10.1038/s43588-024-00741-1.

Using SHAP Post-Model Explainability as a Model-Agnostic Feature Selection Technique

Joshua Gottlieb, Martin Sichali Chunsen, and Alexander Yoo, Krishna Bathula
Seidenberg School of CSIS, University of Pace, New York City, NY, USA
jg05394n@pace.edu, cs83339n@pace.edu, ay71173n@pace.edu, kbathula@pace.edu

Feature selection techniques in data science help reduce dimensionality of datasets, remove sparsity, and eliminate noisy patterns, thereby improving generalization of models to future data. Many feature selection techniques fail to optimize for the predictive power of the model, require iterating through the extensive feature space, or apply only to certain model types [1]. SHAP values focus on capturing the predictive power of each feature within the context of the trained model in order to provide explainability. Because SHAP treats the model as a black box, the SHAP technique is model agnostic. In this paper, we employ SHAP values to capture and rank the important features learned by a model and to use these rankings for feature selection. We build upon prior work using SHAP as a feature selection technique [2] by creating the MAX and SUM selection algorithms, which are designed to align with intuitive feature selection goals without the requirement for costly iterative processes. We validate the effectiveness of our SUM and MAX SHAP selection algorithms by testing their performance with five machine learning models using ten diverse datasets, demonstrating that these SHAP selection strategies produce equivalent or superior model performance and greater feature reduction compared to other state-of-the-art techniques.

Algorithm 1
SUM SHAP Pseudocode

Sort the mean absolute SHAP values I_j in descending order: $I_j \geq I_{j+1}$
Calculate the total sum $S = \sum I_j$
Define a proportion constant r in [0, 1].
Initialize a running sum s and an index $j = 0$.
while $s < r \cdot S$ **do**
  Add I_j to s and increment j
end while
return $[f_0, \ldots, f_j]$, the selected sorted features.

Algorithm 2
MAX SHAP Psuedocode

Sort the mean absolute SHAP values I_j in descending order: $I_j \geq I_{j+1}$
Choose $M = \max(I_j) = I_0$
Define a proportion constant ρ in [0, 1].
Initialize an index $j = 0$.
while $I_j \geq \rho \cdot M$ **do**
  Increment j
end while
return $[f_0, \ldots, f_j]$, the selected sorted features.

References

[1] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014, 40th-year commemorative issue.

[2] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: Assessing SHAP values as feature selection mechanism," in Proc. 33rd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI), Porto de Galinhas, Brazil, 2020, pp. 340–347, doi: 10.1109/SIBGRAPI51738.2020.00053.

Global Misinformation Tracker

Ali Inamdar, Hardik Bhatt, and Krishna Bathula,
Seidenberg School of CSIS, Pace University, New York City, NY, USA
{ai45179n, hb84004n ,kbathula}@pace.edu

The rapid proliferation of misinformation across digital platforms has created significant challenges for information reliability, public trust, and automated content verification. This paper presents the Global Misinformation Tracker (GMT), a scalable system designed to detect, map, and analyze misinformation narratives across domains. GMT integrates heterogeneous data sources from fact-checking repositories, news media, and social platforms into a dynamic knowledge graph that links claims, narratives, sources, and debunks. The framework employs Large Conceptual Models (LCMs) to capture narrative lineage, while a reinforcement learning–based pruning mechanism enhances graph coherence by filtering contradictions and amplifying high-confidence links. Evidence retrieval is facilitated through retrieval-augmented generation (RAG) and vector-based semantic search, enabling verifiable responses with transparent provenance subgraphs. Experimental evaluations on benchmark fact-checking datasets demonstrate improved performance in claim clustering, narrative detection, and misinformation drift tracking compared to baseline methods. By combining explainability, adaptability, and scalability, GMT provides a technically rigorous approach to misinformation monitoring and contributes toward strengthening digital information ecosystems.

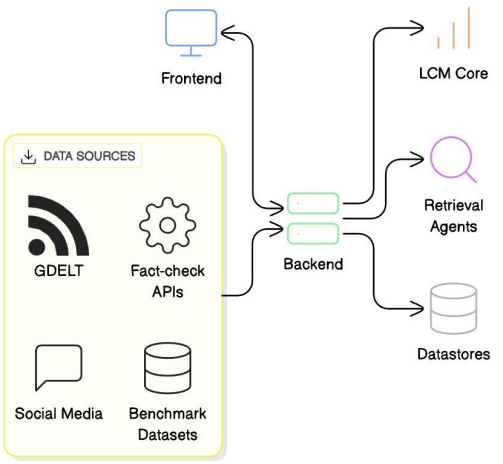


Figure 1: System Architecture of GMT.

References

[1] C. Shao, G. L. Ciampaglia, O. Varol, K. Yang, and F. Menczer, *Tracking and Characterizing the Competition of Fact Checking and Misinformation* IEEE Access, vol. 6, pp. 75340–75351, 2018.

[2] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, *A Comprehensive Survey on Automatic Knowledge Graph Construction* ACM Computing Surveys, vol. 56, no. 4, article 94, 2023.

Fed-CPA: Clustered and Personalized Aggregation for Federated Learning in Heterogeneous Healthcare Networks

Burhan Latief Dar, Farhana Azad, Muhammad Abdullah Zahid, and Krishna Bathula
Seidenberg School Of CSIS Pace University New York, NY, USA
{bd59660n, fa90905n, mz65003n, kbathula}@pace.edu, kbathula@pace.edu

Federated Learning (FL) enables collaborative training of deep learning models across multiple institutions, such as hospitals, without sharing private patient data. This paradigm is crucial for advancing medical diagnostics in areas like digital pathology and radiology. However, the practical application of FL is severely hindered by statistical heterogeneity (Non-IID data), where data distributions vary significantly between clients due to different patient demographics, medical equipment, and specializations. Standard algorithms like FedAvg fail in these settings, as they attempt to force a single global model onto all clients, leading to poor convergence and low personalized accuracy. Existing methods like FedProx offer robustness but still focus on a one-size-fits-all global model, which is suboptimal for specialized clients. To address this critical gap, we propose Fed-CPA (Federated Clustered and Personalized Aggregation), a novel hybrid strategy. Fed-CPA dynamically groups clients into clusters based on the similarity of their model updates, allowing clients with similar data (e.g., "specialist oncology clinics" or "pediatric hospitals") to collaboratively train a shared cluster-specific model. Simultaneously, it uses a personalization component to fine-tune a model head for each client, ensuring high performance on their local data. We validate our approach by simulating a heterogeneous federated network using the public HAM10000 skin lesion dataset. By partitioning the data to reflect realistic clinical scenarios, we demonstrate that Fed-CPA significantly outperforms both FedAvg and FedProx in global model accuracy and, most critically, in personalized diagnostic accuracy for each client.

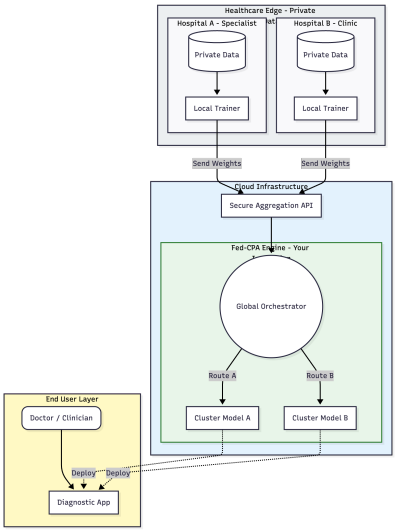


Figure 1: Solution Architecture

Algorithm 1 Fed-CPA: Federated Clustered and Personalized Aggregation

Input: Number of clients K, number of clusters C, rounds T, learning rate η, initial model w_0
Output: Cluster-specific models $\{w^{(1)}, \ldots, w^{(C)}\}$
Server Initialization:
Initialize generic global model $w_{global} \leftarrow w_0$
Initialize cluster models $\{w_0^{(1)}, \ldots, w_0^{(C)}\} \leftarrow w_{global}$
for round $t = 1$ to T **do**
  Client Local Training:
  for each client $k \in \{1, \ldots, K\}$ in parallel **do**
    Receive model $w_t^{(c_k)}$ based on assigned cluster c_k (default to global if $t = 0$)
    Compute local gradients Δw_k on private data D_k
    Send Δw_k to Server
  end for
  Server Clustering:
  if $t == 0$ **then**
    Compute pairwise cosine similarity matrix $S_{ij} = \frac{\Delta w_i \cdot \Delta w_j}{\|\Delta w_i\| \|\Delta w_j\|}$
    Assign clients to clusters $C_1, \ldots, C_C$ based on S
  end if
  Clustered Aggregation:
  for each cluster $c \in \{1, \ldots, C\}$ **do**
    Let S_c be the set of clients in cluster c
    Update cluster model:
    $w_{t+1}^{(c)} \leftarrow w_t^{(c)} - \eta \sum_{k \in S_c} \frac{n_k}{N_c} \Delta w_k$
  end for
  Personalization:
  Clients fine-tune classification head on local data D_k
end for

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

[2] N. Rieke et al., "The future of digital health with federated learning," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[4] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *NeurIPS*, vol. 33, pp. 19545–19560, 2020.

[5] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.

A Regime Aware Deep Ensemble Framework for Adaptive Stock Market Prediction

Raghavi Sendil, Sai Santhosh Channaka, and Krishna Bathula,
Seidenberg School of CSIS, Pace University, New York City, NY, USA
{rs81435n, sc59310n, kbathula}@pace.edu

Forecasting the next day's open, close, high, and low of a stock requires diverse signals like market microstructure, macroeconomic signals, news sentiment, price patterns along with visual pattern intelligence. The traditional model with historical prices will fail during the regime shifts. Our Regime Aware Deep Ensemble framework leverages Hidden Markov Models which combines outputs of three primary layers: (1) The macroeconomic model which uses indicators such as Gross Domestic Product, Consumer Price Index, Purchasing Manager's Index/Institute for Supply Management and interest rates. (2) The microstructure model integrating XGBoost, Long Short-Term Memory based sequential learning, and the Convolutional Neural Network to learn candlesticks and forecast the next day's candle pattern by encoding with geometric features mapping into one of 84 predefined candlestick patterns. (3) A sentiment model using FinBERT embedding with LightGBM. A meta-ensemble with adaptive weighting fuses these predictions. We use Exponential Weighted Moving Average and Generalized AutoRegressive Conditional to estimate volatility while position sizing is carried out by a fractional Kelly criterion. RADE delivers stronger robustness and stability compared to single-source models.

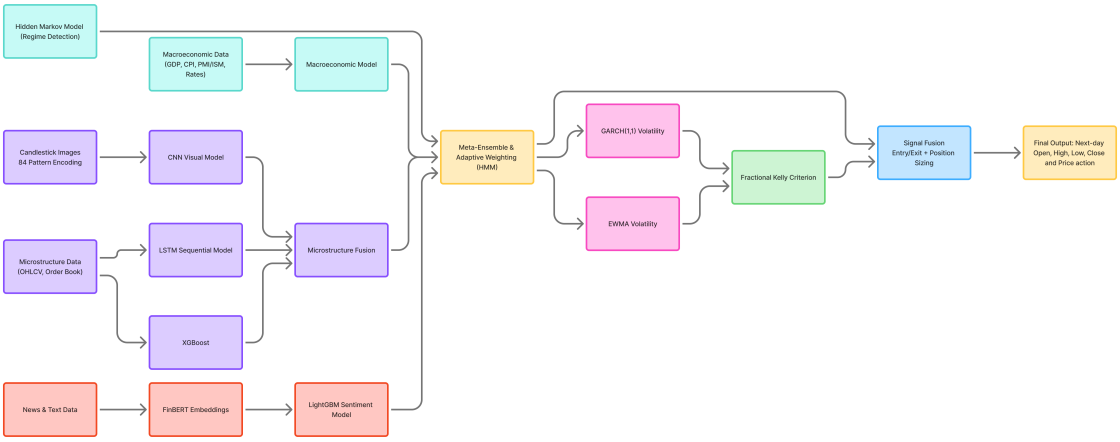


Figure 1: RADE Architectural Flow.

References

[1] Barra S, Carta SM, Corriga A, Podda AS, Recupero DR. 2020. *Deep learning and time series-To-image encoding for financial forecasting.* EEE/CAA Journal of Automatica Sinica 7(3):683-692

[2] Birogul S, Temur G, Kose U. 2020. *YOLO object recognition algorithm and 'buy-sell decision' model over 2D candlestick charts* IEEE Access.

Identifying Determinants of Life Expectancy from World Development Indicators: A Machine Learning Approach

Yekaterina Donegal, Harika Etikela, Abigail Keegan, and Krishna M. Bathula
Seidenberg School of Computer Science and Information Systems
Pace University, New York, NY, USA
{yd16257n, he64457n, ak20905n, kbathula}@pace.edu

Predicting global life expectancy from the World Bank's World Development Indicators (WDI) is challenging due to missing values, high dimensionality, and strong multicollinearity among socioeconomic variables [1, 2]. Prior work shows that the WDI dataset collapses into a small set of latent dimensions, motivating the use of regularized models that remain stable with highly correlated features. This study develops a streamlined supervised learning workflow that engineers a country-level WDI dataset, identifies core predictors, and applies Ridge regression to obtain stable life-expectancy estimates despite multicollinearity. To complement the predictive model, a health-efficiency metric is constructed by regressing life expectancy on GDP per capita and clustering countries based on their resulting performance residuals (the difference between actual and expected LE)[3]. This residual analysis identifies distinct groups of High-efficiency (positive residuals) and Low-efficiency (negative residuals).The results demonstrate that a simple Ridge-based model combined with residual-based efficiency clustering can reveal both predictive structure and meaningful cross-national disparities in health system performance.

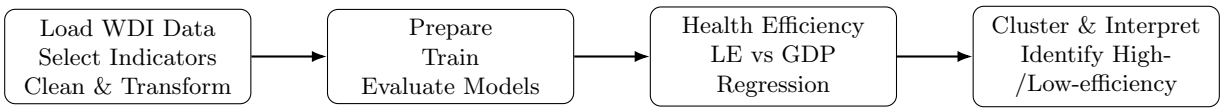


Figure 1: Workflow for Life Expectancy Prediction and Health Efficiency Analysis.

References

[1] A. Kraemer, J. F. Rodrigues, and M. Silveira, *The Low Dimensionality of World Development Indicators*, Social Indicators Research, vol. 152, pp. 113–144, 2020.

[2] S. Hassan, T. Mahmood, and S. Ali, *Machine Learning–Based Analysis of Environmental and Socioeconomic Predictors of Life Expectancy*, Environmental Science and Pollution Research, vol. 29, pp. 51234–51249, 2022.

[3] V. K. Fal'tsman, *Dependence of Life Expectancy of the Population on the Well-Being of the Country (International Statistical Study)*, Studies on Russian Economic Development, vol. 32, no. 2, pp. 194–198, 2021. https://doi.org/10.1134/S1075700721020052

Multimodal Deep Learning for Early Stress Detection Using Micro-Behavioral Signals

Bulty Dolui, Jayashree Johnson, and Krishna Bathula
Seidenberg School of Computer and Information Systems,
Pace University, New York, NY, USA
{bd06753n, jayashree.j, kbathula}@pace.edu

Human overwhelm and stress often rise silently before visible symptoms appear, leading to reduced performance, emotional imbalance, and poor decision-making. This research proposes a non-invasive AI system that predicts early signs of stress before the individual becomes aware of it. By analyzing subtle external signals such as facial micro-expressions, eye-blink rate, voice hesitation, breathing rhythm, posture changes, and reaction-time patterns, the system identifies hidden markers of rising stress. Deep learning models including CNNs, LSTMs, and Transformer-based time-sequence networks are used to detect and forecast stress trends in real time. The aim is to build a proactive support tool that can alert users early, provide personalized recommendations, and help maintain emotional balance. This work has significant potential for improving mental health, workplace productivity, education, driver safety, and human–robot interaction.

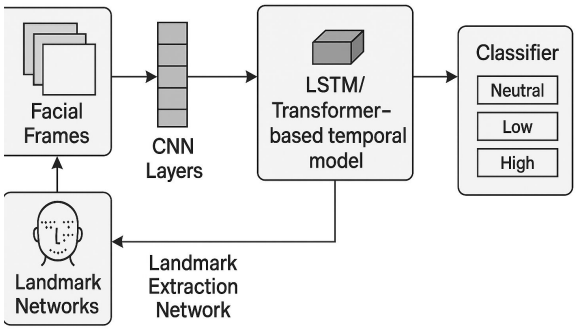


Figure 1: Overall model architecture for multimodal stress detection. [1, 2]

References

[1] Y. Wang, W. Liu, and X. Zhang, "Deep-Learning-Based Stress Recognition with Spatial-Temporal Facial Information," *Sensors*, vol. 21, no. 22, 2021.

[2] A. Sharma and T. Nguyen, "Multimodal Stress Detection Using Facial Landmarks and Biometric Signals," *arXiv preprint* arXiv:2311.03606, 2023.

[3] R. Kumar and S. Saha, "Stress Monitoring Using Machine Learning, IoT and Wearable Sensors," *Sensors*, vol. 20, no. 23, 2020.

Data Driven Disability Accessibility Risk Score for New York City

Sudharsan Vasudevan, Venkata Durga Kavya Bhatta, and Krishna Bathula
Seidenberg School Of Computer Science and Information Systems
Pace University, NY, USA

{sudharsan.vasudevan, venkatadurgakavya.bhatta, kbathula}@pace.edu

This project develops an integrated, multi domain machine learning framework to quantify accessibility risks faced by disabled populations in New York City's boroughs. Public datasets from NYC Open Data and New York State Education Department (NYSED) were curated and engineered to extract domain indicators across four critical sectors such as Education, Transportation, Infrastructure and Employment. Independent ML models were trained for each sector and their predictive performance (accuracy) was used to derive domain level risk coefficients [1]. To align technical output with policy relevance, these coefficients were proportionally weighted using U.S. federal expenditure shares, producing a Budget-Weighted Accessibility Risk Score (BWARS) for each borough[2]. The final scores were stratified into Low, Medium and High risk categories to identify geographic zones with significant accessibility barriers. An interactive interface operationalizes real time risk assessment by allowing users to select borough and age group to generate domain aware accessibility risk profiles. The Data Driven Disability Accessibility Risk Score (DARS) framework supports evidence driven policy design, equitable resource allocation and targeted urban accessibility interventions.

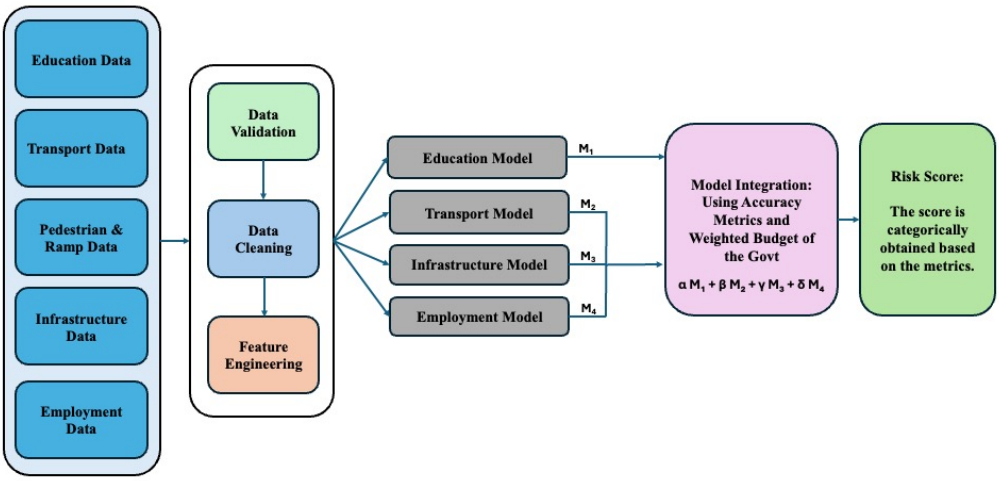


Figure 1: Data Driven Disability Accessibility Risk Score (DARS) for New York City

References

[1] H. Hasan, M. T. Rahman, A. Almogren, and M. A. Rahman, *Machine Learning for Smart Urban Accessibility*, IEEE Access, vol. 10, pp. 12345–12360, 2022.

[2] C.H.Yang, T. Molefyane, and Y.D. Lin, *The Forecasting of a Leading Country's Government Expenditure Using a Recurrent Neural Network with a Gated Recurrent Unit*, ResearchGate, 2023.

An Empirical Evaluation of Classification and Clustering Techniques on Real-World Datasets

Sai Teja Bainapalepu, Ali Inamdar and Krishna Bathula
Seidenberg School of Computer Science and Information Systems,
Pace University, New York, NY, USA
{sb37015n, ai45179n, kbathula}@pace.edu

Machine learning studies often emphasize supervised methods, while the relationship between unsupervised cluster structure and class labels receives less attention [2]. This work provides a unified empirical comparison of both approaches using three Bench marked datasets: Iris, Vertebral Column (VC), and Breast Tissue (BT). We evaluate standard classifiers and measure how clustering aligns with true labels using evaluation metrics. In addition, we analyze how the separability of data influences learning behavior and observe consistent patterns across all datasets. Our findings also show that clustering distortion strongly correlates with supervised accuracy, reinforcing the importance of data geometry [1].

For completeness, our overall result can be summarized as:

$$\downarrow\left(\sum_{i=1}^{n}\|x_i-\mu_{c_i}\|^2\right) \Rightarrow \uparrow \text{classification performance.}$$

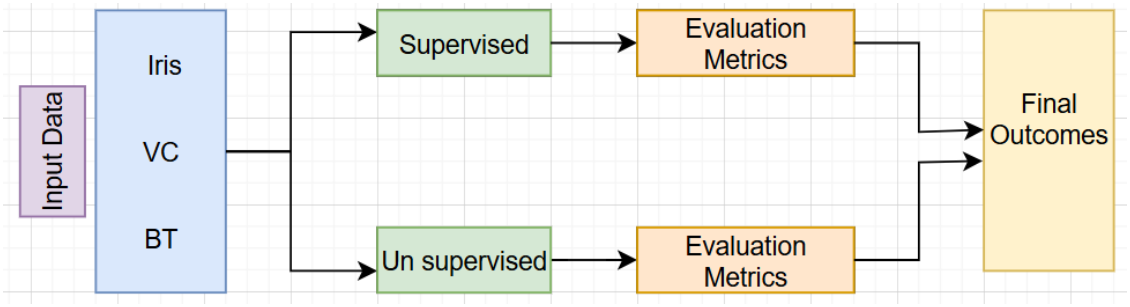


Figure 1: Overall architecture showing supervised and unsupervised learning pipelines leading to final outcome analysis.

References

[1] Y. Yang and A. Blum, *On a Theory of Nonparametric Pairwise Similarity for Clustering: Connecting Clustering to Classification*, NeurIPS, 2014.

[2] M. Ackerman, S. Ben-David, and D. Loker, *Towards Property-Based Classification of Clustering Paradigms*, NeurIPS, 2010.

OncoSummarizer: LLM-Powered Oncology Clinical Note Summarization

Anchal Rai, Viraj Joshi, and Krishna Bathula
Seidenberg School Of Computer Science and Information Systems, Pace University, USA
{ar65667n, vj49006n, kbathula}@pace.edu

The rapid growth of electronic health records (EHRs) in oncology has led to significant challenges for clinicians, who must navigate voluminous, unstructured clinical notes to extract actionable insights. These notes often exceed the context windows of standard language models, complicating efficient information retrieval and decision-making. This paper introduces **OncoSummarizer**, a comprehensive open-source pipeline that uses large language models (LLMs) to address these issues through three core functions: (1) binary classification of oncology-relevant text using fine-tuned transformers, (2) abstractive summarization of cancer-specific notes, and (3) structured extraction of key clinical entities such as cance type, stage, and treatment into JSON format. Drawing inspiration from exclusive systems like those developed by Flatiron Health, our pipeline employs BioClinicalBERT [1] and Longformer [2] for classification on a 50k sample subset of the BlueScrubs dataset, achieving up to 88.77% F1-score. For abstractive summarization, we leverage BART-base [3] fine-tuned on oncology notes, while recent advances in large language model distillation for oncology notes [4] inform our overall approach to efficient, domain-specific summarization and extraction. OncoSummarizer serves as a scalable and privacy-preserving academic tool that reduces clinician workload, facilitates rapid note comprehension, and supports automated documentation and cancer research. Future enhancements include domain-specific fine-tuning and full deployment via Streamlit for interactive use. The pipeline and code will be released publicly to promote further advancements in biomedical NLP.

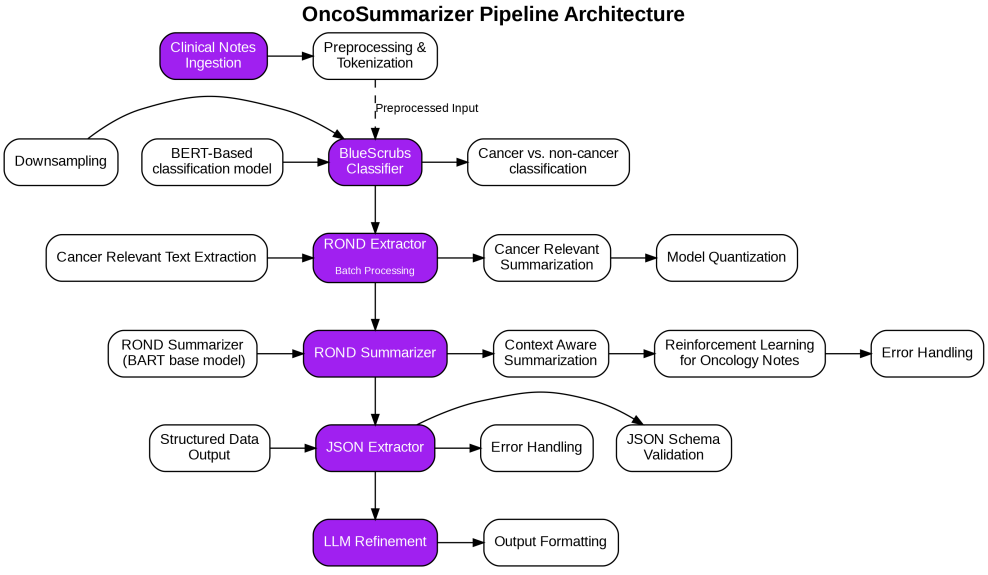


Figure 1: Oncosummarizer Pipeline Architecture

References

[1] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly Available Clinical BERT Embeddings," *Proc. 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019.

[2] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880, 2020.

[4] Y. Zhang et al., "ROND: Rethinking Oncology Note Distillation with Large Language Models,", 2024–2025.

Forecasting New York City Hate Crimes using Time Series Algorithms

Tewan Pothibutt, Vibhav Misra, Nipun Ranchhod Navadia and Krishna Bathula
Seidenberg School Of Computer Science and Information Systems,
Pace University, New York City, NY, USA
{tewan.pothibutt, vibhav.misra, nipunranchhod.navadia, kbathula}@pace.edu

New York is a city known for the title "city that never sleeps". As per the official statiscs from the New York (NYC.gov), 37 percent of New Yorkers are born outside United States and 49 percent speak language other then English. Also it is home to residents who speak over 200 languages. Major portion of these residents and New Yorkers are Asians. This causes hate crimes in the society due to lack of acceptance or any other reasons. NYPD and govt is very active to remove crime from city and they to their best for public safety. To support the authority, residents and tourists we propose the use of time series algorithm (ARIMA, SARIMA, Prophet) to predict the asian hate crime for awareness of people and NYPD in the specific boroughs in next week and month. For this experiment are using NYPD Hate Crime dataset provided by NYPN Police Department which consists various features like Month in which incident occurred, record created date, Patrol Borough Name, County and many more. These features plays a vital role in our study.
Keywords: Time Series, Auto-regressive Integrated Moving Average (ARIMA), Seasonal Auto-regressive Integrated Moving Average (SARIMA), Prophet, Hate Crime, NYPD

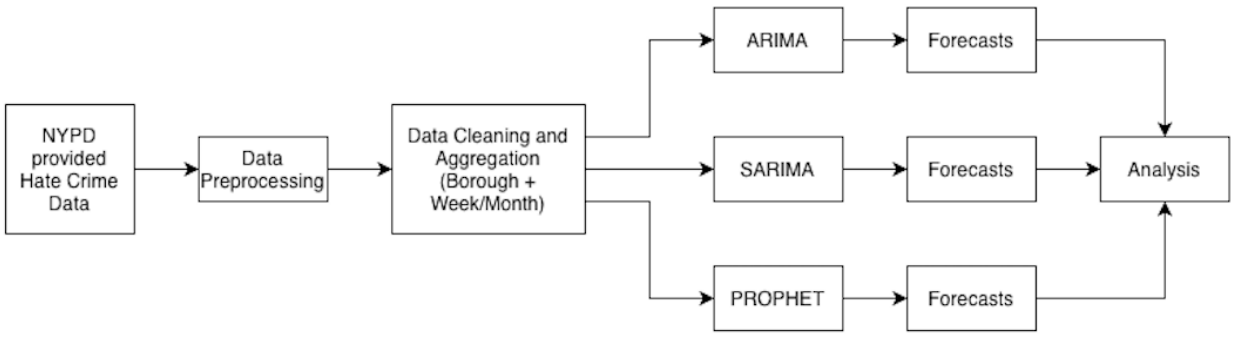


Figure 1: Proposed Experimental Architecture

References

[1] Cho, J., Jeong, J., Seto, C. H.-F. (2025). Community-Level Analysis of Anti-Asian Hate Crime. Crime Delinquency, 0(0). https://doi.org/10.1177/00111287251335012: 1691-1713.

[2] Zhang, Y., Zhang, L., Benton, F. (2022). Hate crimes against asian americans. American journal of criminal justice, 47(3), 441-461.

AI-Based Football (Soccer) Player Market Value Prediction

Phuc Vo
Seidenberg School of Computer Science and Information Systems,
Pace University, New York, NY, USA
pv02523p@pace.edu

The evaluation of the football (soccer) player market is highly subjective and is influenced by reputation, club size, media attention, and transfer speculation. This research investigates whether player market values can be predicted more objectively using performance statistics and machine learning. Using Transfermarkt datasets containing player profiles, performance metrics, and market values, we build a Random Forest regression model and evaluate its ability to estimate player value. The model achieves an R^2 score of 0.411 and a Mean Absolute Error (MAE) of 806,543, indicating that performance-based features explain a significant part of the valuation but do not capture factors driven by prestige and reputation. Feature importance analysis reveals that the current club, minutes played, and position are key predictors. This work provides insights into the fairness and limitations of current market valuations and establishes a foundation for future improvements using richer datasets.

Evaluating the Impact of Imputation, Clustering, and Classification on Credit Default Prediction: A Systematic Pipeline Analysis

Neelima Verma and Sung-Hyuk Cha

Seidenberg School of Computer Science and Information Systems
Pace University, New York, NY, USA
{nv78850n, scha}@pace.edu

Credit default prediction is a critical task in financial risk management, yet most studies focus narrowly on classifier selection while overlooking how upstream pipeline decisions—such as imputation and unsupervised learning—affect predictive performance. This study investigates whether behavioral segmentation obtained from clustering can meaningfully enhance supervised classification accuracy in credit-risk modeling. We construct a controlled 3-stage pipeline incorporating two imputation strategies (Mean, Median), two clustering algorithms (K-Means, K-Medoids), and four classifiers (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting), yielding 16 total pipelines evaluated on the same preprocessed dataset. Results show that clustering plays a substantial role: K-Means consistently produced more informative borrower segments than K-Medoids, leading to stronger downstream classification performance. In contrast, imputation choice had minimal effect. Tree-based ensemble models—particularly Random Forest and Gradient Boosting—performed best, reflecting the nonlinear nature of credit-risk data. Additionally, undersampling outperformed SMOTE in F1-score stability, indicating that simpler balancing methods may generalize better. Overall, the study demonstrates that pairing unsupervised segmentation with nonline

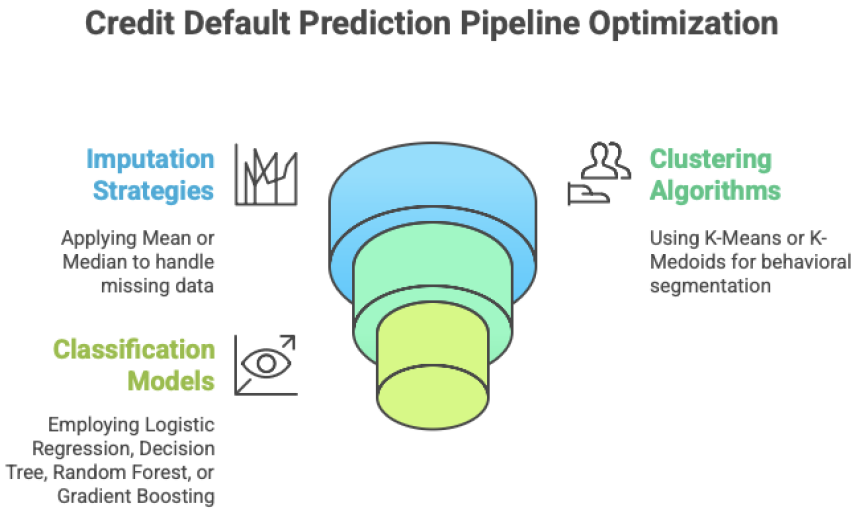


Figure 1: Comparative analysis of pipeline components: imputation methods, clustering algorithms, classifier performance, and class-balancing strategies.

References

[1] Kaggle, *Give Me Some Credit Dataset*, Kaggle Competition, 2011. Available: https://www.kaggle.com/c/GiveMeSomeCredit

[2] S. Lessmann, B. Baesens, H. Seow, and L. Thomas, *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*, European Journal of Operational Research, 247(1): 124–136, 2015.

[3] A. E. Khandani, A. J. Kim, and A. W. Lo, *Consumer credit-risk models via machine-learning algorithms*, Journal of Banking & Finance, 34(11): 2767–2787, 2010.

Identifying Determinants of Life Expectancy from World Development Indicators: A Machine Learning Approach

Fnu Ashutosh, Shivam Jha, and Sung Hyuk Cha
Seidenberg School of Computer Science and Information Systems
Pace University, New York, NY, USA
{an05893n, sj34101n, scha}@pace.edu

Dynamic Time Warping (DTW) is widely used for temporal alignment in physiological signal analysis, yet unconstrained DTW suffers from **pathological warping** in noisy segments - aligning transient artifacts with clinically meaningful morphology (e.g., QRS complexes). Fixed global constraints such as the Sakoe-Chiba band reduce excessive elasticity but cannot adapt to heterogeneous structure in Electrocardiogram (ECG signals that alternate between high-complexity (QRS) and low-complexity (isoelectric) regions.

We present **Entropy-Adaptive Constraint Dynamic Time Warping (EAC-DTW)**, a modified DTW formulation that computes a rolling Shannon entropy profile and maps it through a sigmoid to produce a position-dependent constraint vector. Low-entropy regions receive tight warping limits to suppress singularities; high-entropy regions allow broader alignment flexibility to preserve morphological fidelity.

Using controlled synthetic ECG-like signals (five arrhythmia classes: Normal, LBBB, RBBB, PVC, APC) under three noise conditions (clean, 20 dB, 10 dB SNR), EAC-DTW achieves **79.3% classification accuracy at 10 dB** - improving by **6.0 percentage points** over a fixed 10% Sakoe-Chiba band and outperforming unconstrained DTW in noise robustness. Singularity counts are reduced by **41%** (168 vs 286 for standard DTW), indicating effective mitigation of pathological warping. These results, while promising as proof-of-concept, require clinical validation on real ECG databases for deployment assessment.

Dynamic Time Warping with Multiplicative Combination of Progressive Exponential Weight Penalties

Jinhwi Lee and Sung-Hyuk Cha
Seidenberg School of Computer Science and Information Systems,
Pace University, New York, NY, USA
{jl04999n,scha}@pace.edu

Dynamic Time Warping (DTW) is a widely used metric for comparing time series, providing flexible alignment in the presence of temporal distortions. While classical DTW introduced in [2] allows unrestricted elasticity, amended DTW variants in [1] incorporate penalties to limit excessive warping. However, existing penalties apply constant costs regardless of the extent of elasticity, and previously proposed progressive penalties in [3] only consider extensions on a single sequence. In this study, we introduce two new DTW formulations that apply progressive, exponentially increasing penalties to both horizontal and vertical extensions of the warping path. We propose additive and multiplicative fusion schemes for combining these penalties. Experiments on the FaceFour dataset using 1-NN classification demonstrate that the proposed methods consistently improve performance over classical DTW and constant-penalty DTW. These results highlight the effectiveness and practical value of progressive warping penalties in DTW-based time series analysis.

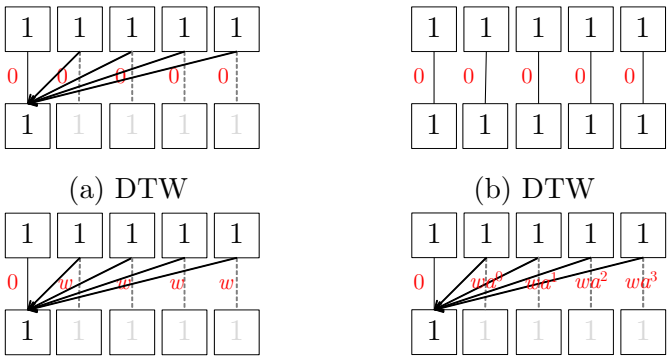


Figure 1: Illustration of Dynamic Time Warping and warping penalties.

References

[1] Matthieu Herrmann and Geoffrey I. Webb. Amercing: An intuitive and effective constraint for dynamic time warping. *Pattern Recognition*, 137:109333, 2023.

[2] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[3] Teja Vuppu, Teryn Cha, and Sung-Hyuk Cha. Enhancing nearest neighbor classification performance through dynamic time warping with progressive constraint. In *Proceedings of 37th International FLAIRS Conference*. The Florida Artificial Intelligence Research Society, 2024.