DATA ANALYTICS 2025

The Fourteenth International Conference on Data Analytics

ISBN: 978-1-68558-293-7

September 28th - October 2nd, 2025

Lisbon, Portugal

DATA ANALYTICS 2025 Editors

Timothy Haas, University of Wisconsin-Milwaukee, USA

Clement Leung (SSE), The Chinese University of Hong Kong, Shenzhen, China

DATA ANALYTICS 2025

Forward

The Fourteenth International Conference on Data Analytics (DATA ANALYTICS 2025), held between September 28th, 2025, and October 2nd, 2025, in Lisbon, Portugal, continued a series of international events on fundamentals in supporting data analytics, special mechanisms, and features of applying principles of data analytics, application-oriented analytics, and target-area analytics.

Processing terabytes to petabytes of data or incorporating non-structural data and multi-structured data sources and types require advanced analytics and data science mechanisms for both raw and partially processed information. Despite considerable advancements in high performance, large storage, and high computation power, there are challenges in identifying, clustering, classifying, and interpreting a large spectrum of information.

We take here the opportunity to warmly thank all the members of the DATA ANALYTICS 2025 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to DATA ANALYTICS 2025. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the DATA ANALYTICS 2025 organizing committee for their help in handling the logistics of this event.

We hope that DATA ANALYTICS 2025 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in the field of data analytics.

DATA ANALYTICS 2025 Chairs

DATA ANALYTICS 2025 Steering Committee
Wolfram Wöß, Institute for Application Oriented Knowledge Processing | Johannes Kepler University, Linz, Austria
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
George Tambouratzis, Institute for Language and Speech Processing, Athena R.C., Greece
Les Sztandera, Thomas Jefferson University, USA
Ivana Semanjski, Ghent University, Belgium
Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands
Joseph G Vella, University of Malta, Malta

DATA ANALYTICS 2025 Publicity Chairs
Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

DATA ANALYTICS 2025
Committee

DATA ANALYTICS 2025 Steering Committee

Wolfram Wöß, Institute for Application Oriented Knowledge Processing | Johannes Kepler University, Linz, Austria
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
George Tambouratzis, Institute for Language and Speech Processing, Athena R.C., Greece
Les Sztandera, Thomas Jefferson University, USA
Ivana Semanjski, Ghent University, Belgium
Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands
Joseph G Vella, University of Malta, Malta

DATA ANALYTICS 2025 Publicity Chairs

Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

DATA ANALYTICS 2025 Technical Program Committee

Arianna Agosto, University of Pavia, Italy
Irfan Ahmed, Virginia Commonwealth University, USA
Madyan Alsenwi, Kyung Hee University, Global Campus, South Korea
Katie Antypas, Lawrence Berkeley National Laboratory, USA
Vincenzo Arceri, University of Parma, Italy
Najet Arous, University of Tunis Manar, Tunisia
Fernando Baptista dos Santos Neves, uConnect Telecom Americas LLC, USA
Abderazek Ben Abdallah, The University of Aizu, Japan
Sadok Ben Yahia, Tallinn University of Technology, Estonia
Soumia Benkrid, Ecole Nationale Supérieure d'Informatique, Algeria
Flavio Bertini, University of Parma, Italy
Nik Bessis, Edge Hill University, UK
Sandjai Bhulai, Vrije Universiteit Amsterdam, Netherlands
Jean-Yves Blaise, UMR CNRS/MC 3495 MAP, Marseille, France
Jan Bohacik, University of Zilina, Slovakia
Vincenzo Bonnici, University of Parma, Italy
Marco Calderisi, Kode srl, Pisa, Italy
Ozgu Can, Ege University, Turkey
Wandercleiton Cardoso, University of Genoa, Italy
Julio Cesar Duarte, Instituto Militar de Engenharia, Rio de Janeiro, Brazil
Junghoon Chae, Oak Ridge National Laboratory, USA
Richard Chbeir, Université de Pau et des Pays de l'Adour (UPPA), France
Daniel B.-W. Chen, Monash University, Australia
Yujing Chen, VMware, USA
Stefano Cirillo, University of Salerno, Italy

Qiang Zhu, University of Michigan - Dearborn, USA
Marc Zöller, USU Software AG / University of Stuttgart, Germany

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Adaptability Challenges in Implementing Big Data Analytics in Tanzanian Small and Medium Enterprises

Peter George Ngugulu
Teesside University International Business School
Middlesbrough, United Kingdom
Email: S3378826@tees.ac.uk

Min Prasad Bhandari
Teesside University International Business School
Middlesbrough, United Kingdom
Email: m.bhandari@tees.ac.uk

Abstract **— This study investigates the adaptability challenges faced by Tanzanian Small and Medium Enterprises (SMEs) in adopting Big Data Analytics (BDA) to enhance operational strategies and strategic decision-making. Grounded in the Technology-Organization-Environment (TOE) framework, the study employed a qualitative approach using semi-structured interviews with SME managers across selected regions. Preliminary insights reveal key challenges to BDA adoption, including infrastructural limitations, skills shortages, financial constraints, and regulatory uncertainties. While the data is limited to a small pilot sample, these findings offer an initial foundation for designing a more comprehensive future study.**

Keywords - Big Data Analytics; SMEs; Tanzania; Technology Adoption; TOE Framework.

I. INTRODUCTION

Small and Medium Enterprises (SMEs) in Tanzania account for over 95% of firms and contribute 35% of Gross Domestic Product [1]. Big Data Analytics (BDA), defined as the process of extracting actionable insights from large, complex datasets [2], can enhance SMEs' operational efficiency and decision-making capabilities. Despite these advantages, Tanzanian SMEs lag in adopting BDA due to infrastructural limitations, financial constraints, and skill shortages [3]. Prior studies have focused on developed economies, or large enterprises, leaving a gap in understanding the Tanzanian SME context [4]. This study investigates the challenges faced by Tanzanian SMEs in adopting BDA and proposes strategies to overcome them.

Tanzania has its economic environment defined by its categorization as a lower-middle-income nation, with an inclined agricultural sector that employs the majority of the workforce and contributes a large amount to GDP. Although this is the case, interest in technology advancement and innovation is increasing in urban cities, where tech hubs and startups are starting to develop [5]. Furthermore, the government of Tanzania has endorsed the utilization of technology as an essential factor [6] in fostering economic activities. It is featured in the development agenda, such as Vision 2025, which has aimed at transforming Tanzania into a semi-industrialized country that focuses on technology and human capabilities.

The paper is structured as follows: In Section II, the literature review examines global and regional perspectives on BDA adoption, identifying gaps specific to Tanzanian SMEs. Section and the research objectives and scope, focusing on adaptability challenges in operational and strategic contexts. Section III presents the core findings on adaptability challenges, divided into two subsections: enhancing operational strategies and strategic decision-making. Section IV discusses the theoretical frameworks and methodology —Technology Acceptance Model (TAM), the Diffusion of Innovations Theory (DOI), and TOE that inform the study's analytical lens. The methodology, including the qualitative approach and data collection techniques. Section V reports the results from interviews with SME managers, illustrating key themes and challenges. Finally, Section VI concludes the paper by outlining future research directions, emphasizing the need to expand the sample size, refine the analytical framework, and explore practical interventions to support BDA adoption in Tanzanian SMEs.

II. LITERATURE REVIEW

BDA has been widely utilized in technologically advanced economies, such as the United States and the United Kingdom, where it has transformed operational strategies, management decisions, and market competitiveness [7][8]. Gandomi et al. [2] described how SMEs in these regions leverage data analytics to improve business processes and customer interactions in considerable detail [2]. Similarly, Asian countries, including China and India, have experienced a growing emphasis on BDA usage in commercial industries, particularly among SMEs, which has fostered innovation and enhanced economic performance [10]. Research conducted in these regions has provided valuable insights into the benefits and challenges associated with the adoption of BDA integration in SMEs.

In contrast, studies focusing specifically on the Tanzanian context continue to show significant gaps [11]. While Mwemezi et al. [11] explored the role of BDA in the banking sector using the TOE framework, limited research exists on how Tanzanian SMEs are adopting BDA to improve operational areas, such as supply chain management, inventory control, and cost efficiency [12].

Although studies like those by Adaga et al. [12] highlight the importance of BDA in decision-making, they do not examine its specific influence on managerial practices within Tanzanian SMEs, particularly considering the unique challenges of the country's economic environment [13].

Even though BDA research has been extensively carried out in different settings worldwide, especially in developed markets, its adoption and effects in emerging markets—particularly among SMEs in Tanzania—remain underexplored. Didas et al. [10], specifically point out a significant gap in the literature on BDA in Tanzania [10]. This is supported by Joubert et al. [13], who emphasize that although BDA is recognized as a potential economic driver, research at the country level in developing nations remains minimal [13].

Additionally, while some studies in Tanzania explore BDA, they do not specifically focus on SMEs. For example, Mwemezi et al. examined BDA in the banking sector [11]. Malero and Seif discussed Hadoop and BDA readiness in Africa generally [14]. Didas et al. explored BDA for managerial large enterprise business-driven decision-making rather than for SMEs [16]. Mkumbo et al. [16], also examined the effect of awareness on BDA adoption readiness in public sector auditing in Tanzania using the TAM [17]. According to Changalima, et al. [17], as cited from Ismail, there is a significant research gap in understanding the technological adoptability capacity in SMEs in Tanzania [18]. In light of this, it becomes evident that while the benefits of BDA are widely discussed, the challenges faced by industries, including SMEs, in adopting BDA—particularly in developing countries—are less frequently covered. Joubert, Murawski, and Bick highlight that there is extensive research on the benefits of BDA, but a significant gap exists in understanding the challenges faced by all African countries in adopting BDA [14]. Similarly, Alalawneh et al. discussed BDA adoption challenges in developing economies and note that, while the advantages are well documented, the challenges are often overlooked [19].

The literature presents a research void on how Tanzanian SMEs are utilizing these technologies to address their unique challenges, including restricted resources, limited infrastructure, and lack of qualified human resources [20]. This exploration addresses this gap by contributing new qualitative data on the acceptance landscape of BDA in Tanzania, extending the work of Ismail et al., who examined low levels of technology absorption among Tanzanian SMEs [12]. Also, according to Ishengoma et al., Tanzanian SMEs—particularly in the manufacturing sector—play a crucial economic role but often lack an appropriate framework for technology adoption tailored to their specific needs [21]. In response to this gap, this study aims to develop a framework that better supports the BDA technology adoption of Tanzanian SMEs, ensuring it aligns with the local conditions and requirements.

Adaptability challenges in the context of BDA refer to the challenges that impede the effective adoption and utilization of BDA technologies [22][23]. The following sections highlight the most prominent adaptability challenges faced by Tanzanian SMEs, categorized by their impact on operational strategies, and strategic decision-making [24]-[27]. The objectives of this work are as follows:

- To identify the adaptability challenges Tanzanian SMEs face in enhancing operational strategies through BDA.
- To explore the adaptability challenges in leveraging BDA for strategic decision-making among Tanzanian SME managers.

III. ADAPTABILITY CHALLENGES IN BDA

A. *Adaptability Challenges in Enhancing Operational Strategies*

BDA has the potential to revolutionize operational strategies by providing real-time insights into supply chain optimization, resource allocation, and cost efficiency. However, Tanzanian SMEs face significant challenges in leveraging BDA for these purposes. Key among these is the lack of affordable and scalable technological solutions, which hinders SMEs from utilizing predictive analytics to optimize their operations [28]. The high costs associated with these technologies and the limited availability of scalable solutions further exacerbate these challenges [29]. Operational inefficiencies often arise from the inability to integrate data from various sources due to fragmented digital infrastructure [25]. For instance, SMEs in Tanzania struggle to adopt tools like predictive maintenance and real-time analytics, which could reduce operational costs and improve efficiency. This gap is compounded by financial constraints, as many SMEs lack the capital to invest in advanced BDA tools and systems [30]. The financial burden of acquiring and maintaining these technologies often leads to underutilization of BDA capabilities, limiting the potential benefits for operational strategies [31].

B. *Adaptability Challenges in Strategic Decision-Making*

BDA facilitates data-driven decision-making by uncovering patterns, trends, and insights critical for formulating effective strategies [32]. However, Tanzanian SMEs often lack the internal expertise required to analyze and interpret complex datasets. This skills gap significantly limits their ability to make informed decisions based on data insights [33]. Additionally, the lack of a data-driven culture and insufficient training further exacerbate these challenges [22]. SMEs face challenges in aligning BDA initiatives with their strategic objectives. Many decision-makers in Tanzanian SMEs remain skeptical about the value of BDA due to limited awareness and understanding of its potential benefits [18]. This skepticism, coupled with the absence of a robust decision-making framework, often leads to underutilization of available data, thereby hindering the

effectiveness of strategic planning [17]. The lack of a clear strategy for integrating BDA into business processes further complicates the adoption and utilization of BDA in strategic decision-making [34].

IV. Theoretical Framework and Methodology

There are several key theories used to understand technology adoption; however, the TAM, DOI, and the TOE framework are more commonly used [35][36].

A. TAM

Developed by Fred Davis in 1989 [36]. TAM focuses on the determinants of technology adoption, emphasizing Perceived Usefulness (PU) and Perceived Ease Of Use (PEOU) as primary factors influencing whether individuals or organizations accept and use new technologies. In the context of Tanzanian SMEs adopting BDA, TAM provides a useful lens to understand how SMEs perceive the benefits of BDA and their apprehension toward complex technology adoption. For example, SMEs in resource-constrained environments, such as Tanzania, may prioritize technologies perceived as easy to use and with visible benefits due to limited budgets, infrastructure, and technical expertise [37]. TAM has been widely utilized in various research contexts to explain user behavior and technology adoption. For instance, Venkatesh and Davis expanded TAM to include external factors such as social influence and facilitating conditions, making it applicable across diverse cultural and organizational settings [38]. Mishrif et al. identified the need for a more nuanced perspective to account for cultural tendencies and infrastructural disparities [39]. Despite its simplicity and wide applicability, TAM's limitations include its narrow focus on individual perceptions and its inability to consider organizational readiness and socio-cultural factors [40].

B. DOI

DOI, introduced by Everett Rogers in 1962 [36], provides a macroscopic view of technology adoption by categorizing adopters into groups: innovators, early adopters, early majority, late majority, and laggards. This segmentation is particularly useful for strategizing the introduction and diffusion of BDA technologies in SMEs. In Tanzanian SMEs, DOI helps identify patterns of adoption behavior, allowing policymakers and technology providers to design targeted interventions for different adopter categories. DOI has been applied in various studies to analyze technology adoption dynamics. For example, during the COVID-19 pandemic, Mishrif et al. highlighted how SMEs rapidly adopted digital technologies to sustain operations, demonstrating the utility of DOI in categorizing adoption behaviors during crises [39].

C. TOE

The TOE framework, introduced by Tornatzky and Fleischer in 1990 [41], offers a comprehensive view of technology adoption by examining three dimensions. The Technological Context includes factors such as compatibility, complexity, and perceived benefits of the technology. The Organizational Context encompasses internal factors like organizational readiness, management support, and employee skills. The Environmental Context involves external factors such as competitive pressure, regulatory environment, and market trends [42]. The TOE framework has been widely utilized in research on technology adoption in SMEs. For instance, Oliveira et al. used the TOE framework to study cloud computing adoption in SMEs, highlighting the significance of external factors like vendor support and industry standards [43].

A qualitative abductive approach was employed, suitable for exploring emerging phenomena [40]. Semi-structured interviews were conducted with five SME managers across cities in Tanzania, namely Dar es Salaam, Arusha, and Dodoma between September and December 2024. Purposive sampling targeted firms in manufacturing, retail, and services. The interviews focused on experiences with data use, perceptions of BDA, and adoption challenges. Data were analyzed thematically using NVivo software to identify recurring patterns aligned with the TOE framework.

V. Results

Theme 1: Operational Efficiency and Data Infrastructure

The ability of BDA to enhance operational efficiency emerged as a dominant theme among the respondents. However, challenges to adopting these capabilities were emphasized. For instance, Respondent 1 highlighted, “We don’t fully understand the benefits and limitations of BDA,” underscoring the lack of awareness and technical knowledge as a key challenge to BDA adoption. This aligns with Babalghaith et al. [34], who found that technical aspects such as complexity and compatibility are significant challenges to BDA adoption in SMEs. Similarly, Willetts et al. [44] identified the lack of expertise and resources as major barriers to effective BDA implementation in SMEs. Respondent 5 echoed similar concerns: “The book used for recording health issues and livestock births got lost, causing a significant loss of information,” which illustrates the absence of proper digital systems for operational data management. This is consistent with findings by Infopulse [29], which discuss how inadequate digital infrastructure can impede the effective implementation of BDA.

Theme 2: Supply Chain and Inventory Management

Overstocking and supply chain challenges were also highlighted. Respondent 2 noted, “When relying on analytics tools, overstocking or, vice versa, stockouts can be avoided due to the ability to forecast demand fluctuations.” This highlights how BDA can mitigate supply chain

inefficiencies by enabling better inventory management and demand forecasting. McKinsey [45] emphasizes that BDA can significantly enhance supply chain decision-making by expanding the dataset for analysis and applying powerful statistical methods to improve inventory management and demand forecasting. Similarly, Cohen [46] discusses how real-time data and machine learning algorithms can revolutionize inventory management, reducing both overstocking and stockouts.

Theme 3: Strategic Decision-Making and Data Collection

Strategic decision-making through market insights was recognized as a key benefit of BDA, but adaptability challenges remain. Respondent 2 observed, "The lack of technical expertise and inadequate data collection from different sources are the main challenges," highlighting the difficulties in generating actionable insights. Respondent 3 stated, "We lack the tools to collect data, such as sensors, to transfer information to computers for recording," which underscores the infrastructural deficits in gathering and utilizing customer data. Additionally, Respondent 4 noted, "Most of our records are done on paper, and we currently only track credit customers, not cash buyers," illustrating the limitations in data collection for creating personalized marketing strategies. Maroufkhani et al. [47] emphasize the importance of top management support and organizational readiness for BDA adoption. This is supported by Babalghaith et al. [34], who found that organizational aspects such as top management support, organizational readiness, and a data-driven culture are crucial for encouraging BDA adoption in SMEs.

Respondent 4 also stressed, "Without proper sales records, we lack performance metrics to rely on," which reflects the limitations in data collection and analysis for effective decision-making. This aligns with the findings of Pingax [48], which discuss the challenges of data collection, including issues of data quality, completeness, and accuracy, that can significantly impact the reliability of insights derived from BDA. Additionally, Willetts et al. [44] highlight that SMEs often struggle with data quality issues, including incomplete, inaccurate, or inconsistent data, which can hinder effective decision-making. The importance of handling missing or incomplete data is emphasized by Cohen [46], who discusses strategies such as data imputation to address gaps in datasets, thereby improving the quality of insights and decision-making processes. The study by McKinsey [45] also underscores the significance of robust data collection and management practices in leveraging BDA for strategic decision-making, noting that reliable data is essential for accurate market insights and effective decision-making.

Theme 4: Organizational Readiness – Resistance from Senior Management

Resistance from leadership was identified as a challenge to BDA adoption. Respondent 5 observed, "The owner, who lacks formal education, is wary of new technologies, fearing that educated employees might take advantage of him." This remark highlights a common challenge in SMEs, where decision-making is often centralized, and leaders may lack the technical literacy to appreciate the potential benefits of advanced technologies. Similarly, Respondent 1 noted, "Our senior managers are hesitant to invest in unfamiliar technologies due to perceived risks." This resistance often stems from a lack of understanding or confidence in the ability of BDA to deliver measurable returns on investment [49]. Leaders who are accustomed to traditional operational practices may view data analytics as an unnecessary complication, further slowing the pace of adoption. To address this resistance, organizations must prioritize leadership engagement and education. Providing senior managers with clear demonstrations of BDA's potential benefits, such as case studies from similar industries, can help alleviate concerns. Additionally, involving leadership in pilot projects and decision-making processes ensures that they feel invested in the technology's success.

VI. Conclusion and Future Work

This pilot study highlights that Tanzanian SMEs face challenges to BDA adoption across the TOE framework. Technologically, inadequate Information Technology infrastructure and limited digital systems hinder readiness; organizationally, weak leadership support and low technical skills remain constraints; and environmentally, unclear policies and limited institutional backing create further challenges. These findings confirm that adoption depends on multiple interrelated factors rather than a single determinant. The research journey offered several lessons. Recruiting SMEs proved difficult due to time pressures and limited awareness of BDA, while language preferences required adaptation of technical concepts. Although participants recognized the value of analytics, financial and expertise gaps restricted engagement. These challenges underline the importance of piloting instruments, building trust, and ensuring cultural sensitivity before scaling up. Future research will expand the sample size across diverse sectors and regions, supported by analysis to prioritize challenges and identify solutions. The next phase will also strengthen the scientific contribution by incorporating tables, graphs, and comparative results to visualize patterns and validate findings. Technical directions include testing low-cost digital tools, targeted training programs, and collaboration with technology providers and industry associations. These steps will refine the TOE framework and generate actionable insights for policymakers and SME leaders.

REFERENCES

[1] URT (United Republic of Tanzania), National Baseline Survey Report for Micro, Small and Medium Enterprises in Tanzania. Dar es Salaam: Ministry of Trade and Industry, 2012.

[2] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, 2011.

[4] London School of Business Administration, Big Data and Strategic Management. London: LSBA Press, 2023.

[5] S. Lyimo, "Digital entrepreneurship and emerging tech hubs in Tanzania," Journal of African Business, vol. 21, no. 4, pp. 450–468, 2020.

[6] N. Kshetri, "The emerging role of big data in key development issues: Opportunities, challenges, and concerns," Big Data & Society, vol. 1, no. 2, 2014.

[7] Zhuang, E. (2025). The impact of big data analytics on business decision-making: A case study. International Journal of Recent Research in Technology. Retrieved from https://ijrrt.com/index.php/ijrrt/article/download/192/67

[8] T. Mucci and C. Stryker, "What is Big Data Analytics?," IBM, Apr. 5, 2024. [Online]. Available: https://www.ibm.com/topics/big-data-analytics. [retrieved: Aug., 2025].

[9] S. Talwar, P. Kaur, A. Dhir, and M. Mäntymäki, "Barriers and benefits of big data analytics adoption in emerging economies: Evidence from SMEs in India and China," Technological Forecasting and Social Change, vol. 173, p. 121033, 2021.

[10] M. Didas and K. Mutajwaa, "Big data analytics adoption challenges in Tanzanian SMEs," African Journal of Information Systems, vol. 15, no. 1, pp. 45–62, 2023.

[11] A. Mwemezi and H. Mandari, "The role of Big Data Analytics in the banking sector of Tanzania: A TOE framework perspective," Journal of African Business, vol. 25, no. 2, pp. 112–130, 2024.

[12] E. M. Adaga, J. Okello, and A. Nyaribo, "Big Data Analytics and decision-making in SMEs: Evidence from Tanzania," Tanzania Journal of Business Studies, vol. 19, no. 1, pp. 88–104, 2024.

[13] P. Joubert, M. Nkosi, and M. Shongwe, "Big data as an economic driver: Challenges and prospects in developing countries," Journal of Development Informatics, vol. 9, no. 1, pp. 77–92, 2023.

[14] B. Malero and S. Seif, "Hadoop and big data readiness in Africa: Opportunities and challenges," African Journal of Information Systems, vol. 5, no. 2, pp. 66–78, 2013.

[15] M. D. Didas, E. Chali, and M. Ndege, "Big data analytics and business-driven decision making: Managerial implications in Tanzania," Journal of African Business Research, vol. 16, no. 1, pp. 15–32, 2024.

[16] M. Mkumbo and P. John, "The effect of awareness on big data adoption readiness in public sector auditing in Tanzania: TAM model perspective," Tanzania Journal of Auditing and Finance, vol. 12, no. 2, pp. 44–60, 2023.

[17] A. Changalima, I. Ismail, and J. Amani, "Technological adoptability capacity in Tanzanian SMEs: An assessment of research gaps," East African Journal of Technology and Innovation, vol. 3, no. 1, pp. 101–117, 2025.

[18] L. Alalawneh and A. Alkhatib, "Challenges of big data analytics adoption in developing economies," International Journal of Information Management, vol. 54, p. 102142, 2020.

[19] S. Kabanda and I. Brown, "A structuration analysis of SMEs' adoption of ICTs in developing countries," Electronic Journal of Information Systems in Developing Countries, vol. 82, no. 1, pp. 1–24, 2017.

[20] E. Ishengoma and M. John, "Technology adoption frameworks for SMEs in the Tanzanian manufacturing sector," African Journal of Management Studies, vol. 10, no. 2, pp. 205–223, 2024.

[21] P. Gamage, "Challenges of big data analytics adoption: A developing country perspective," International Journal of Emerging Technology and Advanced Engineering, vol. 10, no. 6, pp. 22–29, 2020.

[22] Infopulse, "Top big data challenges and solutions for enterprises," 2023. [Online]. Available: https://www.infopulse.com/blog/top-big-data-challenges. [retrieved: June., 2024].

[23] S. Hamisi, "Challenges and opportunities of Tanzanian SMEs in adopting technology," Journal of African Business Research, vol. 7, no. 2, pp. 88–102, 2011.

[24] A. Majeed, M. Yusuf, and M. Saleh, "Digital infrastructure fragmentation and analytics adoption in emerging economies," Global Journal of Information Systems, vol. 15, no. 1, pp. 45–63, 2024.

[25] B. Ponera and S. Kyumana, "Barriers to big data analytics adoption in Tanzanian enterprises," African Journal of Technology and Innovation, vol. 6, no. 1, pp. 89–105, 2024.

[26] F. Ndahani, J. Chuma, and R. Mvula, "Factors impeding big data utilization in Tanzanian SMEs," Journal of East African Business, vol. 12, no. 1, pp. 35–50, 2024.

[27] J. Nikundiwe, "Affordability and scalability of big data solutions for SMEs in Tanzania," Tanzania Journal of Technology and Business, vol. 8, no. 2, pp. 110–123, 2022.

[28] World Economic Forum, The Global Outlook on Big Data Infrastructure and Adoption. World Economic Forum Report, 2025. [Online]. Available: https://www.weforum.org/reports/big-data-infrastructure. [retrieved: Aug., 2025].

[29] A. Shah, R. Khan, and M. Patel, "Financial barriers to big data adoption in SMEs," International Journal of Business Analytics, vol. 7, no. 3, pp. 14–27, 2020.

[30] A. Babalghaith and A. Aljarallah, "The cost of big data adoption in small businesses: A systematic review," Journal of Business Technology Management, vol. 13, no. 1, pp. 75–90, 2024.

[31] S. Chatterjee, N. P. Rana, K. Tamilmani, S. K. Sharma, and Y. K. Dwivedi, "Big data analytics for decision-making: A review and research agenda," Journal of Business Research, vol. 157, p. 113621, 2023.

[32] M. Falahat, Y. Y. Lee, T. Ramayah, P. Soto-Acosta, and C. Lee, "The impact of big data analytics on strategic decision-making capability in SMEs," Journal of Small Business Management, vol. 61, no. 1, pp. 135–153, 2023.

[33] A. Nasrollahi, M. Salehi, and S. Azizi, "Barriers to big data analytics adoption: Evidence from developing countries," Information Systems Frontiers, vol. 23, no. 6, pp. 1493–1512, 2021.

[34] P. Maroufkhani, M. Iranmanesh, and M. Ghobakhloo, "Big data analytics adoption in SMEs: The role of top management support and organizational readiness," Journal of Enterprise Information Management, vol. 36, no. 2, pp. 512–530, 2023.

[35] T. Kika, "The application of technology adoption models in emerging economies: A review," African Journal of Information Systems, vol. 10, no. 1, pp. 54–65, 2018.

[36] H. O. Awa, O. U. Ojiabo, and B. C. Emecheta, "Integrating TAM, TOE, and DOI frameworks in SMEs adoption of e-commerce in Nigeria," Journal of Internet Commerce, vol. 14, no. 2, pp. 1–27, 2015.

[37] U. Thathsarani and W. Jianguo, “Factors affecting SMEs’ intention to adopt big data analytics: Evidence from resource-constrained environments,” Journal of Small Business and Enterprise Development, vol. 29, no. 5, pp. 873–891, 2022.

[38] V. Venkatesh and F. D. Davis, “A theoretical extension of the technology acceptance model: Four longitudinal field studies,” Management Science, vol. 46, no. 2, pp. 186–204, 2000.

[39] A. Mishrif and A. Khan, “SMEs’ technology adoption strategies during COVID-19: A diffusion of innovations perspective,” Journal of Business Research, vol. 152, pp. 102–114, 2023.

[40] A. Bryman, Social Research Methods, 4th ed. Oxford: Oxford University Press, 2012.

[41] L. G. Tornatzky and M. Fleischer, The Processes of Technological Innovation. Lexington Books, 1990.

[42] I. Arpaci, Y. Yardimci Cetin, and O. Turetken, “Organizational adoption of information technologies: A literature review within the TOE framework,” Procedia Technology, vol. 1, pp. 118–124, 2012.

[43] T. Oliveira and M. F. Martins, “Literature review of information technology adoption models at firm level,” Electronic Journal of Information Systems Evaluation, vol. 14, no. 1, pp. 110–121, 2011.

[44] J. Willetts, M. Atkins, and C. Stanier, “Overcoming barriers to big data analytics adoption in SMEs: A resource-based view,” International Journal of Information Management, vol. 58, p. 102402, 2021.

[45] McKinsey & Company, The Big Data Opportunity: Supply Chain Management and Analytics. McKinsey Global Institute, 2023. [Online]. Available: https://www.mckinsey.com. [retrieved: Aug., 2025].

[46] M. C. Cohen, “Inventory management and big data: Balancing costs, risks, and benefits,” Production and Operations Management, vol. 24, no. 6, pp. 1074–1085, 2015.

[47] P. Maroufkhani, M. Iranmanesh, and M. Ghobakhloo, “Big data analytics adoption in SMEs: The role of top management support and organizational readiness,” Journal of Enterprise Information Management, vol. 36, no. 2, pp. 512–530, 2023.

[48] Pingax IT Solutions, “Top 7 data quality challenges and how to overcome them,” Pingax Insights, 2023. [Online]. Available: https://www.pingax.com. [retrieved: Jan., 2024].

[49] D. Lopez, Z. Wang, and M. Rahman, “Leadership resistance in digital transformation: Barriers to adopting data-driven technologies in SMEs,” Journal of Small Business Strategy, vol. 33, no. 1, pp. 45–60, 2023.

Parametric Optimization and Intelligent CAD Automation for Bicycle Frame Design via Multi Agent

Chon Kit Chan[1], Wen-Yi Yang[2], Jin H. Huang[3]
[1] Department of Aerospace and Systems Engineering
[2] Master's Program of Electro acoustics
[3] Department of Mechanical and Computer Aided Engineering
Feng Chia University
Taichung, Taiwan (R.O.C.)
e-mails: p1130484@o365.fcu.edu.tw

Abstract—**This research presents an intelligent design platform integrating Large Language Models (LLMs), Retrieval Augmented Generation (RAG), a multi-agent system, and the Model Context Protocol (MCP). It addresses the inefficiencies, high costs, and heavy reliance on manual expertise in traditional bicycle design workflows, which constrain both innovation and responsiveness to market demands. This challenge is especially pressing as the bicycle industry faces increasing demands for customization, small-batch production, and rapid product development. While prior approaches such as Machine Learning (ML) and data-driven optimization have shown promise, they remain confined to isolated tasks and lack a unified, end-to-end framework. The proposed system leverages these technologies to generate context-aware design recommendations tailored to user intent and automates Computer Aided Design (CAD) model generation, thereby substantially reducing development time and manual workload. As a proof of concept, the platform is first applied to rim design the component with the richest dataset before scaling to other components and full frame development. Experimental results demonstrate that the system can reduce design cycles from weeks to hours, providing higher efficiency, lower costs, and improved accuracy of design recommendations. The goal is to accelerate intelligent, flexible, and automated design workflows for the next generation of bicycle products.**

Keywords—Bicycle rim design; large language model; retrieval-augmented generation; multi-agent system; intelligent CAD automation.

I. INTRODUCTION

Taiwan assumes a critical position in the global bicycle industry, renowned for its leading manufacturing technologies. However, the evolving market demand for highly customized, small batch, and diversified products presents persistent difficulties to Taiwan's bicycle industry, particularly in structural innovation and aesthetic design. Traditional design workflows are predominantly dependent on the experience and manual efforts of engineers. This includes extensive data collection (e.g., rim profiles, aerodynamic characteristics, weight, stiffness) and iterative design modifications. Such processes are not only time-consuming, often requiring over a month to complete an initial design, but also incur high Research and Development (R&D) costs and involve a large volume of repetitive tasks.

The current bicycle design workflow faces several critical challenges. Firstly, low efficiency prolongs design cycles and hampers responsiveness to market shifts. Secondly, a strong dependence on manual labor and iterative testing inflates R&D costs. Thirdly, significant time spent by designers on tedious data collection and model adjustments limits their capacity for creative innovation. Lastly, excessive reliance on senior engineers' experience complicates systematic knowledge management, application, and transfer.

However, existing approaches remain insufficient to overcome these challenges. Traditional CAD and engineering workflows are heavily manual, making them slow, costly, and difficult to scale. While data-driven and ML methods have recently been introduced, most prior efforts are restricted to narrow tasks such as parameter prediction or dataset construction. These approaches lack a unified framework that connects data analysis, knowledge retrieval, and CAD automation, limiting their applicability to complex, multi-component bicycle designs.

To overcome these limitations, this study introduces an innovative intelligent design platform integrating LLMs, RAG, a multi-agent system, and the MCP. The platform's core objective is twofold: to intelligently generate design parameters and recommendations aligned with user defined intent, and to automate the generation of CAD models, thereby alleviating the workload of manual drafting. Implementation of these features is anticipated to significantly shorten product development cycles, reduce R&D costs, and enhance both design efficiency and quality.

The remainder of this paper is organized as follows: Section 2 provides a review of relevant literature and technologies; Section 3 provides an in-depth explanation of the proposed AI system architecture and research methodology; Section 4 showcases the experimental results and offers analysis in Results and Discussion; and Section 5 summarizes the key contributions of this study and discusses directions for future research.

II. RELATED WORK

In recent years, Artificial Intelligence (AI) technologies, particularly LLMs, have made significant advancements across diverse domains and are increasingly demonstrating potential in complex engineering design and automation workflows. LLMs have shown practical value in specialized fields, such as medicine [1], and their capacity for natural language comprehension and generation has opened new avenues for enhancing human machine interaction, for instance, in human robot communication [2]. Academic

research further indicates that LLMs can continuously refine their performance via self-improvement mechanisms [3] and have been successfully applied in areas like personalized recommendation systems [4].

Within intelligent design automation, especially for product specific applications like bicycle design, data-driven approaches have garnered considerable attention. For example, the BIKED dataset, developed by Regenwetter et al., provides valuable data resources for bicycle design and has facilitated the establishment of ML benchmarks and applications in this domain [5] [6]. These studies underscore the feasibility of employing data and ML for analyzing and optimizing design parameters.

This study utilizes an SQL Agent that interacts with databases via LLMs, aligning with current advancements in Text to SQL technologies. Numerous academic efforts have concentrated on augmenting the SQL generation capabilities of LLMs. These include integrating domain knowledge to improve accuracy (e.g., the Knowledge to SQL approach) [7], conducting systematic reviews of LLM based Text to SQL methodologies [8], and employing strategies like model fusion to enhance overall performance [9] and accuracy on real world relational databases [10]. These advancements provide a solid foundation for developing more natural and intelligent data querying and analysis interfaces.

The design platform proposed herein integrates RAG to bolster the model's domain-specific knowledge and query response accuracy, particularly concerning bicycle rim design. Furthermore, the platform adopts a Multi-Agent System architecture. This architecture is designed to decompose complex design tasks and delegate them to specialized agents with distinct functionalities, thereby improving the efficiency and modularity of the overall design workflow.

While the aforementioned technologies (LLMs, RAG, Multi-Agent Systems, MCP) have individually shown progress in specific applications, their integrated application to parametric optimization and CAD automation—particularly for complex components like bicycle rims and frames requiring multifaceted engineering considerations —remains a challenging yet promising research avenue. This study aims to offer an innovative solution to the limitations of traditional design workflows through the development of a unified intelligent design platform.

III. Methodology and AI System Architecture

This section introduces the methodology and overall system architecture of the proposed intelligent design platform. By integrating LLMs, RAG, a Multi-Agent System, and the MCP into a unified framework, the approach aims to overcome the inefficiencies and limitations of traditional bicycle design workflows. The subsections describe the system design, the construction of each agent, and the user interface.

A. Overall AI System Architecture

The AI-based bicycle rim design system developed in this study features a backend architecture as illustrated in Figure 1. The system is designed to handle user tasks, such as: “I would like to design a bicycle rim with low aerodynamic drag, low weight, and high stiffness. Please provide design recommendations and generate a CAD drawing.”

At the core of the system is the Main Agent, responsible for receiving and interpreting user tasks, decomposing them, and distributing them to specialized sub agents within the Multi-Agent framework. It then consolidates their responses to generate a complete reply for the user. Considering that bicycle design often involves parameter analysis, textual reference consultation, and CAD drafting, this study introduces three key sub agents. The first is the SQL Agent, which automatically generates SQL queries to retrieve parameter data from the database; if the required data is not available, it leverages ML techniques to predict the parameters. The second is the RAG Agent, which enriches the system’s domain knowledge in bicycle rim design to ensure the quality and accuracy of generated responses. The third is the CAD Agent, which uses the verified parameters—proposed collaboratively by the SQL and RAG Agents—to generate both 2D profiles and 3D models.

This study also employs prompt engineering to establish a horizontal communication mechanism among agents, enabling effective information flow between them. This approach enhances the overall reliability of user responses and helps mitigate the risk of potential “data silos.”

In this study, Claude was adopted as the primary LLM based on response quality, computational efficiency, and practical usability. This choice ensured stable and reliable performance across different tasks without requiring additional reconciliation between multiple models.

B. Training Data

The AI training data used in this study is diverse, supporting various system functionalities. The first component is structured data, which includes the correlation between geometric dimensions and corresponding performance metrics (such as aerodynamic drag, weight, and stiffness) of specific bicycle rim models (e.g., Model A). This study utilizes a dataset generated by the Taiwan Bicycle Research Center (CHC) through Finite Element Method (FEM) simulations. The second component is textual data. To enhance the model’s knowledge base, this study incorporates RAG techniques. The related training texts are primarily sourced from extensive online technical materials—such as design manuals, technical reports, and research articles—collected using the research capabilities of LLMs (e.g., OpenAI ChatGPT, XAI Grok, and Google Gemini). The third component consists of CAD file data. As one of the study's objectives is to enable automated rim CAD drawing generation, a standardized CAD dataset was constructed to train the AI model in learning the geometric features and drawing logic of rims, which also serves as a basis for verifying the generated outputs.

C. Core AI Technologies

The design platform developed in this study integrates several cutting-edge AI technologies to enable highly automated and intelligent design capabilities. At its core is

the LLM, which is trained on deep learning architectures such as Transformers and excels in natural language understanding and generation. Within the system, the LLM plays a central role in semantic parsing, task reasoning, and user interaction for various intelligent agents. ML, a crucial subfield of AI, builds predictive models through data-driven algorithms; in this system, ML techniques are primarily applied within the SQL Agent, which uses trained regression models to predict parameters when they are unavailable in the database. Intelligent agents are autonomous computational entities capable of perceiving their environment, reasoning, and taking actions to achieve specific goals. The agents in this system combine the semantic capabilities of the LLM with operational tools (e.g., database query modules and CAD control APIs), allowing for multi-step reasoning and automation of complex tasks. The MCP is a system architecture designed for multi-module AI systems. It emphasizes modular encapsulation of various models and functional units, coordinated via a unified command protocol, which enhances the system's manageability and scalability by enabling dynamic integration and flexible execution of tools. Lastly, RAG combines information retrieval with generative modeling by first retrieving relevant documents or text fragments from external knowledge bases and then feeding them to the LLM as context for content generation and question answering. This significantly improves performance in knowledge intensive tasks in terms of accuracy, reliability, and interpretability.

The evaluation in this study was based on standard predictive performance metrics, including R^2, MSE, MAE, and RMSE, reported in the experimental section. A goal specification was also defined, e.g., "generate a rim CAD model with weight < 450 g and stiffness > 120 N/mm." This goal was decomposed into parameter retrieval, property estimation, and CAD generation, handled by the SQL, RAG, and CAD Agents. A goal was considered achieved once all sub-goals met predefined thresholds, with satisfactory levels verified against industry standards. In practice, the CHC industry partner confirmed that the generated results met their design requirements and were considered satisfactory, further validating the achievement criteria.

D. Construction Process and Details of Each Agent

The overall system architecture and workflow of this study are illustrated in Figure 2. After defining the goals and framework of the AI system, each agent was independently constructed and trained. The development of the RAG Agent began with the collection of a large corpus of domain specific texts related to bicycle rim design. These texts were preprocessed by extracting relevant content, segmenting it into appropriately sized chunks, and converting them into vector representations. Subsequently, a retrieval mechanism based on a Faiss vector database was established, and the RAG model was trained and fine-tuned to optimize both retrieval efficiency and generation quality. To guide the reasoning pathway and operational logic of the RAG Agent, carefully crafted prompts were designed. Finally, the trained RAG model and its supporting tools—such as the Faiss search engine and metadata filtering modules—were encapsulated as MCP modules and subjected to multiple rounds of testing and validation.

The construction process of the SQL Agent begins with cleaning and organizing the optimized parameters obtained through Finite Element Method (FEM) simulations and storing them in a structured database. The next step involves developing functionalities that enable the agent to automatically generate SQL query statements and execute database searches. Since the FEM dataset contains only around 500 data entries, ML is employed to predict values that fall outside this range or are missing from the database. To this end, a multi-output regression model is adopted and trained for prediction tasks. Prompts are also designed to guide the SQL Agent's query logic and the conditions under which the prediction model should be triggered. Finally, the SQL query engine, SQL code generator, and ML based prediction executor are encapsulated into an MCP module, and both the query results and prediction outputs are validated for accuracy and reasonableness.

The development of the CAD Agent begins with the preparation of standardized bicycle rim CAD files as training data. The agent is then trained to understand the drawing logic and geometric specifications of CAD models. Utilizing FreeCAD MCP Tools, which include a Python code generation module and a FreeCAD API controller — the agent sends generated Python drawing commands via API to the FreeCAD software for automated generation of 2D and 3D CAD models. During this process, prompts are also designed to guide the CAD Agent in translating input parameters into appropriate drawing instructions. All related functionalities are encapsulated into an MCP module, followed by tests to verify the accuracy and completeness of the generated models.

Once the three core agents are trained and packaged, they are integrated into the Main Agent for system level integration testing and validation. If the system passes the validation, it is ready for deployment; otherwise, the process returns to the corresponding agent's training phase for refinement and optimization.

E. Frontend User Interface (UI)

The frontend User Interface (UI) of the system is designed to provide a convenient and intuitive interaction experience and is implemented using CURSOR, as shown in Figure 3. Users can simply enter their design requirements or questions in natural language through the chat box on the right side of the interface (highlighted as Area 2 in red in Figure 3) to interact with the AI design system. The left side of the interface (Area 1 in red) displays the currently active agents and available MCP tools, enhancing the transparency of system operations.

IV. Results and Discussion

To evaluate the effectiveness of the AI design system, we employed the latest large language model, Claude 4 Sonnet [11], as the core reasoning engine for each agent. A simulated question and answer scenario replicating a

realistic design task was created to assess each agent's response capabilities, the efficiency of horizontal communication between agents, and the overall quality of information integration and final response by the Main Agent. An example test prompt is as follows: "I would like to design a bicycle rim with moderate aerodynamic drag, stiffness, and weight. Please recommend an optimal set of dimensional parameters, analyze the reasonableness of these parameters based on the design manual, then draw the 2D and 3D models in FreeCAD, and finally provide a summary and design recommendations." The following sections present the functionality and analysis of each agent:

A. SQL Agent Response

In response to the user's query, "I am currently looking to design a bicycle rim with moderate aerodynamic drag, moderate stiffness, and moderate weight. Please recommend an optimal set of size parameters," the SQL Agent first queried the internal database. As shown in the red box in Figure 4, the agent recommended the following optimal parameters: outer width of 34 mm, inner width of 19 mm, and total length/depth of 79 mm. An initial assessment suggests that this response is reasonable — for instance, greater aerodynamic requirements often correspond to increased rim depth. However, the actual validity of these parameters requires further analysis by the RAG Agent.

B. RAG Agent Response

Following the parameters provided by the SQL Agent, the user then asked, "Please analyze the reasonableness of this set of parameters." The response from the RAG Agent, as shown in the red box in Figure 5, indicates that it successfully communicated horizontally with the SQL Agent, retrieved the parameters, and performed an analysis based on the embedded knowledge from the bicycle rim design manual. In this analysis, the RAG Agent offered suggestions regarding the feasibility of each dimension and the types of bicycles for which the parameters would be appropriate. Due to space limitations in the paper, only a portion of the key content is presented here. The full analysis includes multiple aspects, and users can further inquire about how to refine the current parameters for improved design suitability.

C. CAD Agent Response

This study trained the CAD Agent using 2D profile drawings, such as the example training data shown in Figure 6. After the SQL and RAG Agents confirm the design parameters, the CAD Agent receives instructions to perform modeling. As shown in the red box in Figure 7, the CAD Agent confirms the modeling dimensions—outer width of 34 mm, inner width of 19 mm, and total depth of 79 mm—consistent with the previously recommended values. Figure 8 highlights the 2D profile and 3D solid model successfully generated by the CAD Agent in FreeCAD, precisely matching the input parameters. These results demonstrate that the CAD Agent can execute instructions effectively, significantly reducing the time and manual effort required for CAD drafting by the user.

D. Main Agent Response

Figure 9 presents the response from the Main Agent of the AI-based bicycle rim design system. It consolidates the processes and analyses of all the aforementioned agents and provides concrete design recommendations. This system enables users to efficiently achieve low volume, high variation design objectives. Moreover, the agents are equipped with real time online search capabilities, enhancing their domain knowledge and adaptability.

E. ML Multi-Output Regression Model Training Result

Figure 10 displays the training results of the ML multi-output regression model. In this study, a parameter prediction model was trained using the following variables: bicycle rim cross section outer width (mm), inner width (mm), total length (mm), FEM simulation rigidity (N/m), FEM simulation weight (g), and FEM simulation wind resistance (gf). Performance metrics were used as input features, while dimensional parameters served as outputs. This allows the model to predict a broader range of parameter combinations based on user requirements, rather than being limited to the 500 FEM simulated samples. Prior to training, the linear relationships between dimensions and performance indicators were verified. As shown in Figure 10, the model performed well overall. However, one parameter exhibited signs of overfitting, with an MSE of 0 and an R^2 of 1, which occurred because the parameter in question was a constant value.

F. ML Multi-Output Regression Model Training Result

This study adopted Claude 4 [12] as the primary LLM due to its stable reasoning performance, consistent response quality, and efficient integration with external tools. Compared with alternatives such as ChatGPT or Gemini, Claude 4 delivered more reliable outcomes for multi-agent collaboration and CAD automation tasks. For real-world deployment, the proposed agentic framework is technically feasible, as its modular, MCP-based design enables seamless integration with existing CAD and database systems. Nonetheless, additional factors including data security, system latency, and compatibility with large-scale production workflows must be addressed. This collaboration with CHC indicated that the system outputs aligned with industry expectations, suggesting strong potential for adoption in actual production environments.

V. Conclusion and Future Work

This study successfully developed an intelligent CAD automation design platform that integrates LLMs, RAG, a Multi-Agent System, and the MCP, specifically aimed at optimizing bicycle rim parameters and reengineering the traditional design workflow. By incorporating the SQL Agent for parameter querying and prediction, the RAG Agent for domain knowledge enhancement, and the CAD Agent for 2D and 3D modeling, the system effectively addresses the high time and labor costs associated with manual data collection and repetitive drafting in conventional design processes. Experimental results

demonstrate that the platform can reduce design time from several weeks to just a few hours, significantly improving both efficiency and the accuracy of design recommendations.

Looking ahead, the intelligent design platform proposed in this study holds significant potential for further development and expansion. First, its current capabilities in rim design can be gradually extended to other critical bicycle components, such as frames and forks, ultimately enabling intelligent design assistance at the full vehicle level. Second, the integration of materials databases and related analytical modules is envisioned, allowing the system to consider the impact of material selection on performance and cost during the early stages of design. Third, deeper integration with advanced Computer Aided Engineering software could automate more complex simulations involving structural mechanics, aerodynamics, and beyond. Fourth, continuous improvements to the user interface and interactive experience are necessary to develop more intuitive and intelligent human machine collaboration modes, thereby lowering the barrier to use. The ultimate goal is to explore higher level automated product development paradigms, applying AI to conceptual design, engineering design, manufacturing planning, and even product lifecycle management—further advancing smart manufacturing and digital transformation.

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Council of Taiwan under Grant NSTC 113-2221-E-035-042. Additional support was provided by the Cycling & Health Tech Industry R&D Center (CHC) and an unrestricted gift from Google.org. The authors gratefully acknowledge this support. Translation from the original language into English has been aided by an automatic tool.

REFERENCES

[1] I. L. Alberts, L. Mercolli, T. Pyka, G. Prenosil, K. Shi, A. Rominger, et al., "Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?," Eur. J. Nucl. Med. Mol. Imaging, vol. 50, no. 6, pp. 1549–1552, 2023.

[2] C. Y. Kim, C. P. Lee, and B. Mutlu, "Understanding large language model (LLM) powered human robot interaction," in Proc. 2024 ACM/IEEE Int. Conf. Human Robot Interaction, 2024, pp. 371-380.

[3] J. Huang, S. S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, et al., "Large language models can self improve," arXiv preprint arXiv:2210.11610, 2022.

[4] H. Lyu, S. Jiang, H. Zeng, Y. Xia, Q. Wang, S. Zhang, et al., "LLM Rec: Personalized recommendation via prompting large language models," arXiv preprint arXiv:2307.15780, 2023.

[5] L. Regenwetter, B. Curry, and F. Ahmed, "BIKED: A dataset and machine learning benchmarks for data driven bicycle design," in Proc. Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., vol. 85383, 2021, p. V03AT03A019.

[6] L. Regenwetter, B. Curry, and F. Ahmed, "BIKED: A dataset for computational bicycle design with machine learning benchmarks," J. Mech. Des., vol. 144, no. 3, p. 031706, 2022.

[7] Z. Hong, Z. Yuan, H. Chen, Q. Zhang, F. Huang, and X. Huang, "Knowledge to SQL: Enhancing SQL generation with data expert LLM," arXiv preprint arXiv:2402.11517, 2024.

[8] Z. Hong, Z. Yuan, Q. Zhang, H. Chen, J. Dong, F. Huang, and X. Huang, "Next generation database interfaces: A survey of LLM based text to SQL," arXiv preprint arXiv:2406.08426, 2024.

[9] T. Zhang, C. Chen, C. Liao, J. Wang, X. Zhao, H. Yu, et al., "SQLfuse: Enhancing text to SQL performance through comprehensive LLM synergy," arXiv preprint arXiv:2407.14568, 2024.

[10] G. M. Coelho, E. R. Nascimento, Y. T. Izquierdo, G. M. García, L. Feijó, M. Lemos, et al., "Improving the accuracy of text to SQL tools based on large language models for real world relational databases," in Int. Conf. Database Expert Syst. Appl., Cham, Switzerland: Springer, 2024, pp. 93–107.

[11] Anthropic, Claude Opus 4 & Claude Sonnet 4 System Card, May 2025. [Online]. Available:https://www.anthropic.com/claude-4-system-card

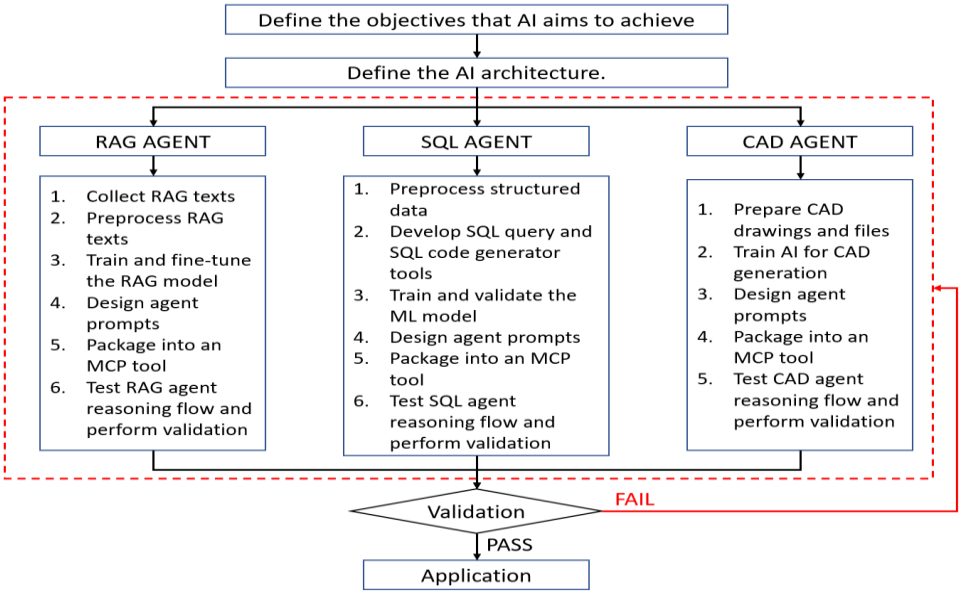

Figure 1. Research Architecture and Workflow.

FRONT-END

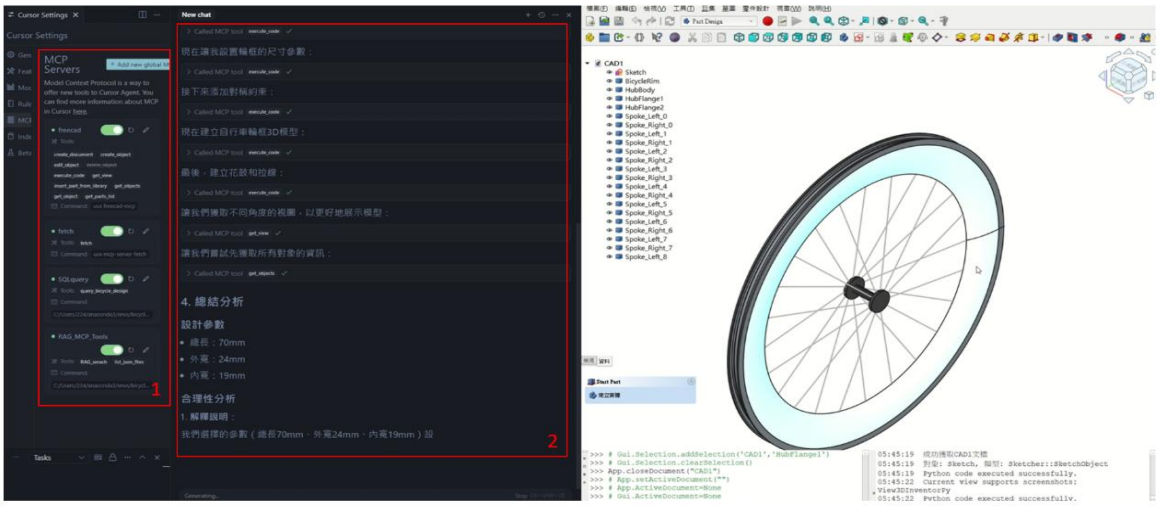

Figure 2. Frontend UI of the AI-Based Bicycle Rim Design System.

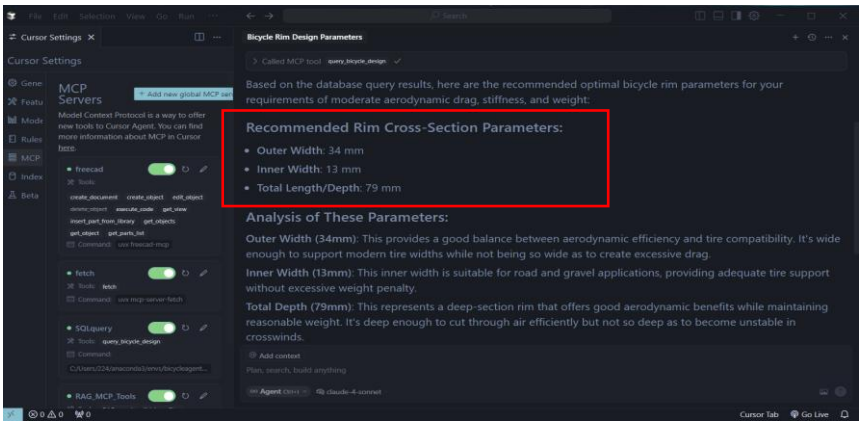

Figure 3. SQL Agent Response.

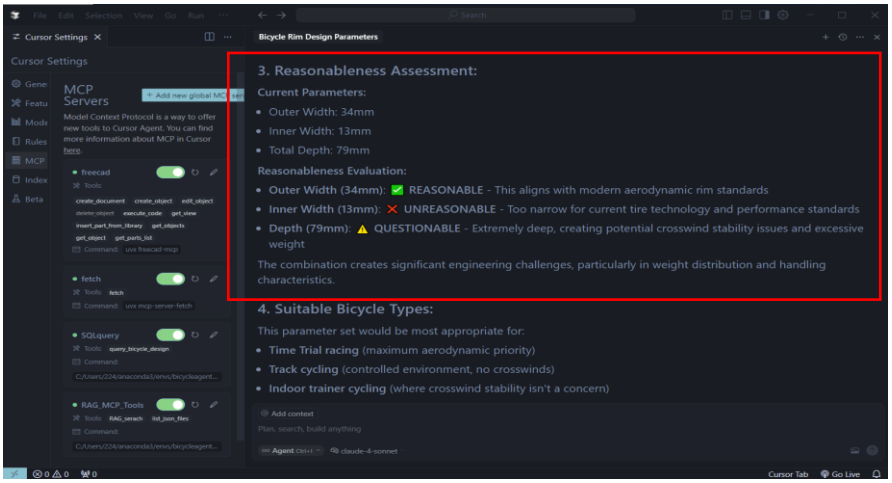

Figure 4. RAG Agent Response.

Figure 5. 2D Drawing of CAD Training Data.

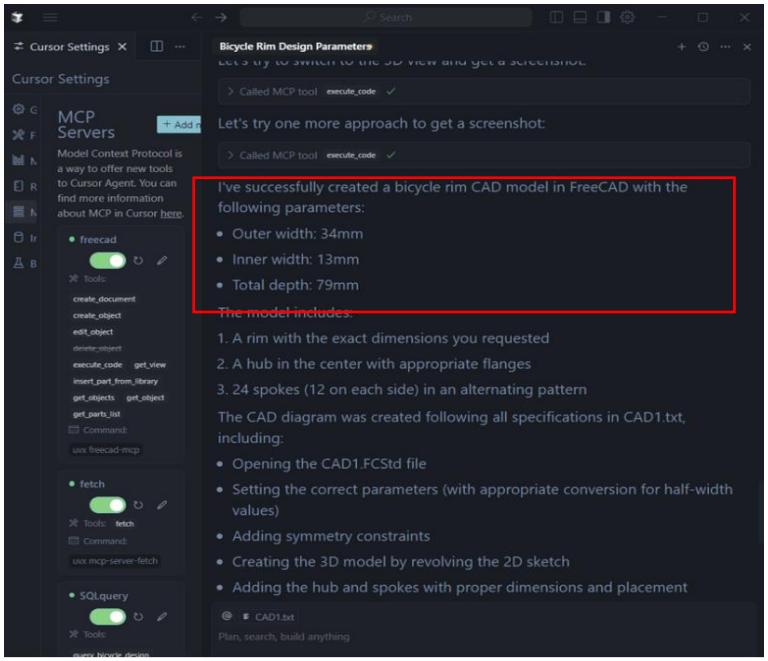

Figure 6. CAD Agent Response.

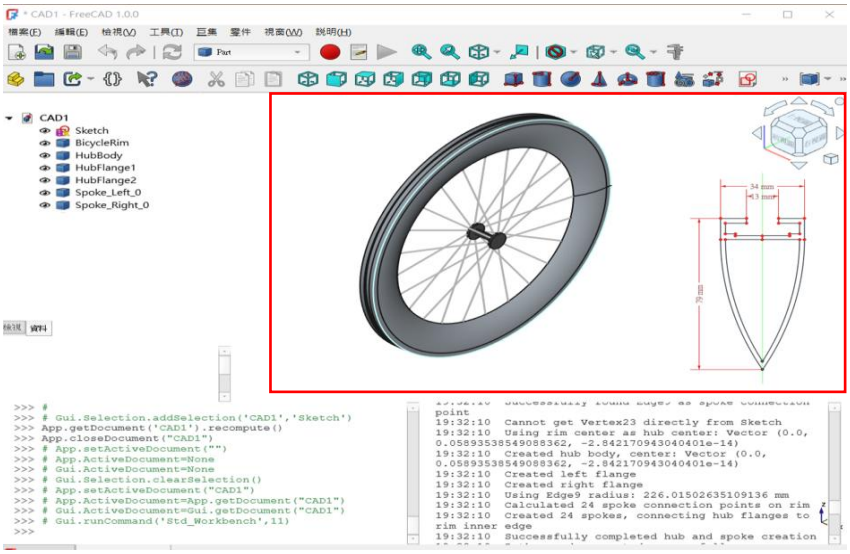

Figure 7. CAD Agent 2D and 3D Modeling.

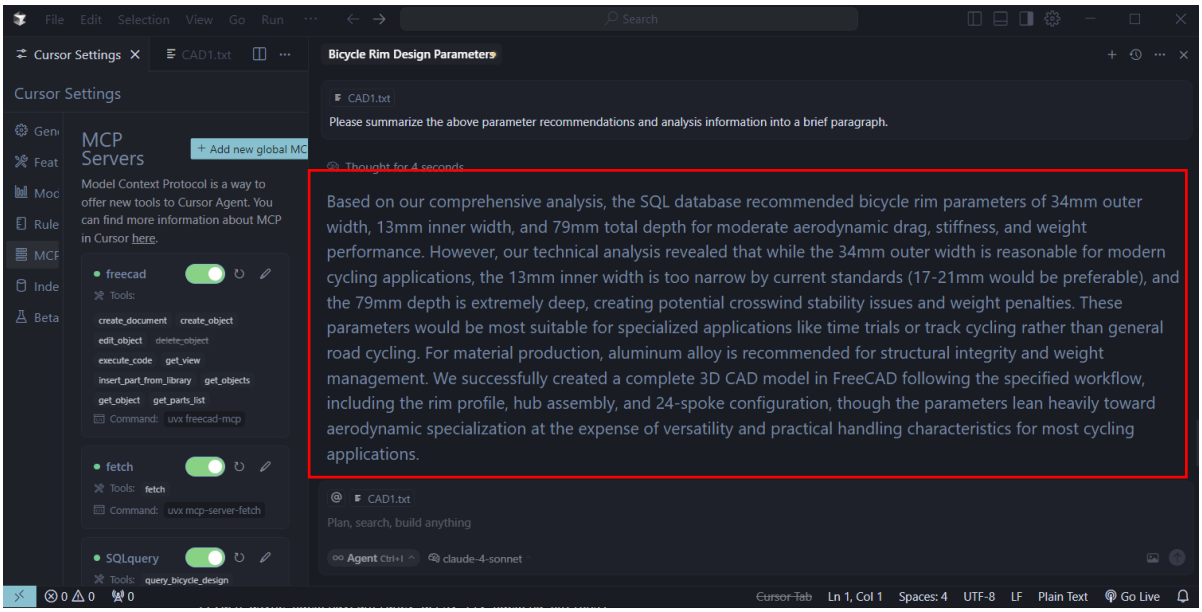

Figure 8. Main Agent Response of the AI-Based Bicycle Rim Design System.

```
bicycle rim cross section outerwidth(mm) - Mean Squared Error: 0.5412679611650484, R²: 0.9703689907840284
bicycle rim cross section innerwidth(mm) - Mean Squared Error: 0.0, R²: 1.0
bicycle rim cross section totallength(mm) - Mean Squared Error: 7.299524271844662, R²: 0.9759634249223411
```

Figure 9. Training Results of the ML Multi-Output Regression Model.

Trigger Injection via Clustering for Backdoor Attacks on Heterogeneous Graphs

Honglin Gao, Lan Zhao, Gaoxi Xiao
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
e-mail: HONGLIN001@e.ntu.edu.sg, zhao0468@e.ntu.edu.sg, egxxiao@ntu.edu.sg

***Abstract*—Heterogeneous graph neural networks have achieved remarkable success in modeling multi-relational data. However, the risks associated with backdoor attack have largely gone unexplored. In this paper, we present a new structure-based backdoor attack method for heterogeneous graph neural networks. Our method uses a set of designed trigger nodes in the graph connected to semantically related parts of the graph using clustering-based trigger node selection. Triggering nodes cause the model to misclassify certain target nodes as an attacker-specified class while still keeping a high accuracy on the clean data. Preliminary experiments on publicly available benchmark datasets show that our proposed backdoor attack is effective and stealthy. This shows that there is a clear need for security awareness in heterogeneous graph learning.**

***Keywords-heterogeneous graph; backdoor attack*.**

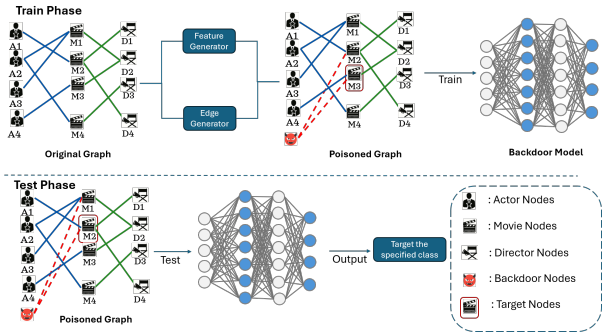

Figure 1. Backdoor process.

I. INTRODUCTION

Heterogeneous Graph Neural Networks (HGNNs) have quickly become prominent for leveraging multi-typed relational data, such as through recommendation [1], social analysis [2] and financial intelligence applications [3]. While HGNNs have gained much success in applications, there has been a lack of investigation into their security. Just like their homogeneous counterparts, HGNNs are susceptible to backdoor attacks; intentional corruption of a model such that the model will perform incorrectly when using a certain trigger. Backdoor attacks are serious attacks that have been overlooked by many researchers.

Unlike traditional adversarial attacks, backdoor attacks implant a hidden pattern during training that causes abnormal responses to specific triggers. While such attacks can target various graph-based tasks, this work focuses on the classification setting, where at inference time the model behaves normally on clean data but misclassifies target nodes into attacker-specified classes when the input contains the trigger.

Numerous notable approaches have been proposed for homogeneous graph backdoor attack, such as Unnoticeable Graph Backdoor Attack (UGBA) [4] and Clean-label Graph Backdoor Attack (CGBA)[5]. Specifically, UGBA employs structure-level triggers by optimizing the triggering structure's topological similarity with benign substructures, with the goal of minimizing the visibility of perturbation, and avoiding structural detection. Alternatively, CGBA utilizes a clean-label approach by injecting feature-based triggers into nodes belonging to the target class, without any modifications to the labels or graph structure. In general, both methods assume same node types with homogeneous edge semantics, and lack modeling mechanisms to adequately represent semantic constraints in heterogeneous graphs, such that their overall attack effectiveness cannot be fully realized.

Our method achieves target-specific misclassification without compromising the overall performance of the model. This is accomplished by adding carefully selected trigger nodes and forming edges that are consistent with the types of the key regions. The preliminary results suggest that these assaults are highly effective and difficult to detect, raising concerns over the security of HGNN-based systems. This method has several problems because it needs the capacity to change the network topology and a complete understanding of graph schemas, which makes it less useful in black-box situations.

The remainder of this extended abstract is organized as follows. Section II introduces the proposed backdoor attack framework designed for heterogeneous graphs. Section III presents experimental results on the IMDB dataset(a comprehensive online databases of movies, TV shows, actors, and production crew information), which shows the relationships between movies, actors and directors, to validate the effectiveness and stealthiness of the attack. Section IV provides a comparative discussion with existing methods, and Section V concludes the paper with future directions.

II. METHODS

We propose Heterogeneous Backdoor Attack (HeteroBA), a structure-manipulating backdoor attack framework tailored for heterogeneous graphs. The core idea is to insert a node of a type that can legally connect to the target node type, and generate semantically coherent features for it using a Feature Generator. To provide coherence with the relational constraints of heterogeneous graphs, an Edge Generator links this trigger node to other present nodes in a type-consistent manner. This

TABLE I. BACKDOOR ATTACK EFFECTIVENESS ON IMDB DATASET

Dataset	Victim Model	Class	Trigger	ASR				CAD			
				HeteroBA-C	HeteroBA-R	CGBA	UGBA	HeteroBA-C	HeteroBA-R	CGBA	UGBA
IMDB	HAN	0	director	**0.9953**	0.6791	0.5618	0.2087	0.0307	0.0265	**0.0037**	0.0364
		1		**0.9984**	0.8458	0.4523	0.2991	-0.0031	-0.0094	**-0.0119**	0.0037
		2		**1.0000**	0.9003	0.4992	0.3582	0.0068	**-0.0068**	0.0010	0.0067
	HGT	0	director	**0.8473**	0.7975	0.4851	0.5109	0.0036	0.0021	**-0.0104**	0.0291
		1		**0.9299**	0.8878	0.4147	0.7757	0.0182	−0.0146	**0.0130**	0.0026
		2		**0.8894**	0.8193	0.4523	0.6807	0.0026	**-0.0099**	-0.0015	0.0182
	SimpleHGN	0	director	**0.9533**	0.7679	0.3881	0.8443	-0.0047	0.0015	**-0.0244**	0.0005
		1		0.9502	0.9486	0.3850	**0.9595**	0.0047	0.0052	**-0.0130**	0.0291
		2		**0.9720**	0.8255	0.3474	0.9330	**-0.0052**	−0.0166	0.0156	0.0078

enables the injected node to propagate misleading information to the target node while maintaining high stealthiness.

The overall attack pipeline is illustrated in Figure 1. The Feature Generator and Edge Generator work together to insert a crafted trigger node (e.g., the red node) into the graph, connecting it to semantically relevant regions. During training, the trigger is embedded into the model without degrading clean performance. At test time, when the same structural pattern reappears and connects to a target node, it activates the backdoor behavior, causing the target to be misclassified into the attacker-specified class.

In order to provide stealth for the implanted trigger nodes, the Feature Generator gives them feature vectors with properties close to those of benign nodes of the same type. This is done by modeling the feature distribution in relation to the trigger node type using Kernel Density Estimation (KDE) [6]. KDE is a non-parametric method of estimating the probability density function of a random variable. In our case, it captures the empirical feature distribution of clean nodes from the target class.

For a given set of clean nodes of a certain type (e.g., "author" nodes in an academic network), we apply Kernel Density Estimation (KDE) to get a smoothed estimate of their feature space. We then sample new feature vectors for trigger nodes from this estimated distribution. This method guarantees that the generated features are statistically indistinguishable from those of valid nodes, thereby rendering it hard to identify trigger nodes via feature-based anomaly detection methods.

To enhance the influence of the trigger and improve its stealthiness within the graph structure, the Edge Generator in HeteroBA adopts a clustering-based strategy to determine how the trigger node is connected. Specifically, we first identify a subset of nodes that are legally allowed to connect to the target node type according to the schema of the heterogeneous graph. We then perform clustering within this subset based on node feature information, dividing the candidates into several semantically coherent and structurally compact regions.

After clustering, we select some influential nodes and connect the trigger node to them. Not only does this approach ensure legitimate edge types, but it also inserts the trigger into a meaningful local context. Compared to random, clustering-driven edge selection improves the attack effectiveness while preserving the graph's overall structure, making the backdoor harder to detect.

III. RESULTS

We evaluate our method on the widely used IMDB dataset, which is a heterogeneous graph composed of three node types: movies, directors, and actors, with edges representing semantic relations such as directed-by and acted-in. In our attack setting, director nodes are injected as trigger nodes to manipulate the classification results of movie nodes. A visual comparison of the graph before and after trigger injection is presented in Figure 1.

To measure attack effectiveness, we employ two conventional evaluation metrics. The Attack Success Rate (ASR) is the ratio of poisoned target nodes that are misclassified into an attacker-chosen label at inference. The Clean Accuracy Drop (CAD) shows how much test accuracy goes down on clean data, which shows how stealthy the attack is [4].

As shown in Table I, under the HAN model [7], our proposed method HeteroBA-C (which uses clustering-based edge injection) achieves over 99% ASR across all target classes, significantly outperforming baselines such as CGBA and UGBA. The CAD, on the other hand, stays within ±0.01, which means that clean data is not affected much.

To further validate the effectiveness of the clustering-based edge injection strategy, we compare HeteroBA-C with a variant called HeteroBA-R, in which the injected trigger node connects to randomly selected legal-type nodes instead of semantically coherent clusters. Table I shows that HeteroBA-R has a much lower ASR, while CAD is similar to HeteroBA-C. This contrast shows that clustering-based structural placement greatly improves the effectiveness of attacks without sacrificing stealthiness.

IV. DISCUSSION

The comparison of the IMDB dataset indicates that HeteroBA works far more effectively when dealing with graphs that have multiple types of nodes and connections. By incorporating type-compatible trigger nodes and semantically consistent edges, our approach attains better attack efficacy with minimal disturbance to accurate predictions.

In terms of computational cost, the dominant overhead of HeteroBA lies in the clustering-based auxiliary node selection. Let p denote the number of target nodes and n_{aux} the number of auxiliary nodes. For each target node, HeteroBA performs a clustering operation with complexity $O(n_{\text{aux}} \log n_{\text{aux}})$, resulting in an overall time complexity of $O(p \cdot n_{\text{aux}} \log n_{\text{aux}})$.

Other steps such as feature sampling and edge insertion incur negligible cost. This demonstrates that HeteroBA balances both attack performance and computational scalability, making it feasible for practical use in real-world heterogeneous graphs.

CGBA, on the other hand, only foucuses on feature-level modification by finding the most discriminative feature dimension and using it as a trigger in the poisoned nodes. This method is simple, it lacks structural adaptability. More importantly, in real-world applications such as social networks or recommendation systems, directly altering node features (e.g., modifying user profiles or item attributes) is often impractical or easily detectable. HeteroBA's method of adding additional nodes or edges, on the other hand, is both achievable and undetected, making it easier to fit into existing graph structures.

UGBA, on the other hand, employes a bi-level optimization strategy: The inner loop increases the classification confidence of the poisoned nodes, while the outer loop uses cosine distance to make sure that the features are similar at the feature level. This design works well in homogeneous environments, but it does not quite capture the complex semantics of different node types and relationships found in heterogeneous graphs. As a result, its triggers lack contextual compatibility, reducing both effectiveness and stealth.

The consistently low CAD values across multiple classes and models confirm that HeteroBA is not too noticeable. These findings clearly demonstrate that HGNNs are particularly susceptible to backdoor attacks that exploit their structural awareness. They stress the urgent need to create targeted protection mechanisms that are tailored to the specific semantics and realistic structures present in these models.

V. Conclusion and Future Work

In this work, we propose HeteroBA, a heterogeneous graph neural network backdoor attack framework that perturbs structure specifically crafted for the task. By co-designing feature and edge generators according to the graph schema, HeteroBA is able to inject semantically plausible triggers that cause targeted misclassification with minimal negative impact on clean data. Experiments on the IMDB dataset validate its high attack success rate and remarkable stealth capabilities over baselines with demonstrated performance.

For future work, We intend to expand HeteroBA toa broader range of heterogeneous graph datasets in different domain, including academic networks (e.g., DBLP-a computer science bibliography website) [8], e-commerce networks (e.g., Amazon) [9], and extensive bibliographic graphs (e.g., OAG-a large knowledge graph unifying two billion-scale academic graphs) [10]. To solve scalability problems in such large graphs, we will look into more efficient versions of our approach, including clustering with sampling or mini-batch KDE-based feature generation, to reduce the amount of computing resources needed without lowering performance. We also intend to evaluate our method under more diverse victim models, including Graph Attention Networks (GAT) based [11] and transformer-based heterogeneous Graph Neural Networks (GNNs) [12]. In addition, we aim to explore adaptive defense mechanisms capable of detecting or neutralizing structure-aware backdoors in heterogeneous settings.

Acknowledgement

The authors gratefully acknowledge the partial support of the Ministry of Education, Singapore, under the research grant RG10/23.

References

[1] A. Salamat, X. Luo, and A. Jafari, "Heterographrec: A heterogeneous graph-based neural networks for social recommendations", *Knowledge-Based Systems*, vol. 217, p. 106 817, 2021.

[2] D. Singh and A. Verma, "An overview of heterogeneous social network analysis", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 15, no. 2, e70028, 2025.

[3] S. Xiang, D. Cheng, C. Shang, Y. Zhang, and Y. Liang, "Temporal and heterogeneous graph neural network for financial time series prediction", in *Proceedings of the 31st ACM international conference on information & knowledge management*, 2022, pp. 3584–3593.

[4] E. Dai, M. Lin, X. Zhang, and S. Wang, "Unnoticeable backdoor attacks on graph neural networks", in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2263–2273.

[5] X. Xing, M. Xu, Y. Bai, and D. Yang, "A clean-label graph backdoor attack method in node classification task", *Knowledge-Based Systems*, vol. 304, p. 112 433, 2024.

[6] Y.-C. Chen, "A tutorial on kernel density estimation and recent advances", *Biostatistics & Epidemiology*, vol. 1, no. 1, pp. 161–187, 2017.

[7] X. Wang *et al.*, "Heterogeneous graph attention network", in *The world wide web conference*, 2019, pp. 2022–2032.

[8] X. Fu, J. Zhang, Z. Meng, and I. King, "Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding", in *Proceedings of the web conference 2020*, 2020, pp. 2331–2341.

[9] X. He *et al.*, "Lightgcn: Simplifying and powering graph convolution network for recommendation", in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 639–648.

[10] W. Hu *et al.*, "Open graph benchmark: Datasets for machine learning on graphs", *Advances in neural information processing systems*, vol. 33, pp. 22 118–22 133, 2020.

[11] P. Veličković *et al.*, "Graph attention networks", *arXiv preprint arXiv:1710.10903*, 2017.

[12] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model", *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

Achieving Near Real-Time Data Freshness in Fraud Detection: An HTAP Approach

Matteo G. Giorgino
Department of Computer Information Systems
Faculty of ICT, University of Malta
Msida, Malta
e-mail: matteo.giorgino.18@um.edu.mt

Joseph G. Vella
Department of Computer Information Systems
Faculty of ICT, University of Malta
Msida, Malta
e-mail: joseph.g.vella@um.edu.mt

***Abstract*—The rise of complex financial fraud in banking demands sophisticated detection solutions capable of near real-time operations, delivering rapid responses on fresh data. Traditional architectures that connect Online Transactional Processing (OLTP) and Online Analytical Processing (OLAP) through Extract-Transform-Load (ETL) pipelines often fail to satisfy these requirements, particularly when both data consistency and rapid response times are critical. This paper examines how Hybrid Transactional/Analytical Processing (HTAP) architectures can address these limitations by consolidating transactional and analytical workloads within a single system. To assess HTAP's suitability for Fraud Detection in near real-time scenarios, the paper employs HyBench, a benchmarking framework that measures data freshness in centralised HTAP systems, augmented with a custom-made external harness. This setup allows for systematic scenario exploration and detailed performance tracking under realistic banking workloads. Across 48 hours, the evaluation executes 96 parameterised runs at multiple data volumes, with database configurations optimised for the available hardware. Results indicate that an HTAP platform can sustain continuous access to fresh data, achieving sub-20 ms freshness, even under mixed OLTP and OLAP loads, while maintaining high transactional throughput. Although there are efficiency trade-offs compared to standalone OLTP or OLAP deployments, proper system configuration and tuning prove critical for balancing performance and freshness. Furthermore, the flexible benchmarking harness developed here enables practitioners to define custom metrics and integrate additional processing logic into the pipeline, extending beyond HyBench's capabilities.**

Keywords–HTAP System; Near Real-Time Fraud Detection; Database Architecture; Data Freshness; DBMS Benchmarking.

I. INTRODUCTION

Effective Fraud Detection in the financial sector hinges on the availability of relevant, timely and high-quality data [1]. By scrutinising transaction records for distinctive patterns and anomalies, modern systems can distinguish legitimate activities from suspicious or malicious behaviour. The efficacy of these systems depends on computational capabilities that guarantee near real-time data freshness, enabling rapid anomaly detection, and swift pattern recognition across vast datasets. The ultimate objective is to classify each transaction accurately and immediately, preventing fraud without impeding business operations.

In practice, Fraud Detection platforms must ingest data from a variety of heterogeneous sources. Each source contributes unique attributes that, when integrated, offer a comprehensive view of user behaviour. However, differences in data formats, transmission frequencies and endpoint capabilities can introduce synchronisation challenges. Data arrival is often asynchronous, leading to temporary mismatches between sources that must be reconciled to preserve analytical accuracy. Ensuring consistent data quality and low-latency access under such dynamic conditions requires robust strategies for data ingestion, harmonisation and error handling. Without adaptive synchronisation mechanisms, systems risk suffering latency spikes, incomplete data views or compliance gaps that undermine detection performance.

Many organisations have adopted architectures that separate Online Transactional Processing (OLTP) from Online Analytical Processing (OLAP), which are linked by Extract-Transform-Load (ETL) workflows. While this paradigm maintains a clear boundary between transactional consistency and analytical throughput, it introduces inherent latency and operational overheads. Batch-oriented ETL processes can struggle to satisfy the stringent low-latency requirements of fraud detection and prevents OLTP systems from directly leveraging the most recent analytical insights during transaction authorisation.

This paper explores Hybrid Transactional/Analytical Processing (HTAP) systems as a unified architecture capable of addressing these limitations. HTAP platforms merge transactional and analytical workloads, offering continuous, low-latency access to up-to-date data without the need for discrete ETL cycles. By running analytics directly on live transactional feeds, HTAP can enhance both responsiveness and detection accuracy, reducing the window of vulnerability to fraud. We present a detailed evaluation of HTAP performance in a banking context, measuring data freshness, throughput and anomaly-detection latency under realistic workload scenarios. To evaluate HTAP's effectiveness for near real-time Fraud Detection, we select a benchmarking suite that defines and measures freshness on a centralised HTAP system and extends it with a custom-built external harness. This setup allows for systematic scenario enumeration and detailed performance tracking across realistic banking workloads. Our findings demonstrate that, with careful configuration and tuning, HTAP systems can sustain sub-20 ms data freshness while processing high transaction volumes, outperforming OLTP/OLAP deployments in near real-time fraud detection.

Despite their advantages, HTAP systems also come with limitations. Running transactional and analytical workloads concurrently can cause resource contention, especially under heavy loads. Tuning performance requires expertise, and some platforms may lack support for strong consistency or scalability. Upgrading from traditional architectures also involves significant cost and complexity.

This paper begins with Section I, which introduces the problem and motivation. Section II sets out the research aims and objectives. Section III provides background on fraud detection in banking, system development, the role of Database Management Systems (DBMSs), database architectures, HTAP technologies, and benchmarking. Section IV details the schema, benchmarking suite, and experimental setup. Section V presents the results, focusing on the main Key Performance Indicators (KPIs). Finally, Section VI summarises the findings and outlines directions for future work.

II. AIMS AND OBJECTIVES

The Fraud Detection System (FDS) aims to enable near real-time classification of fraudulent transactions by ensuring data freshness, and facilitating ad hoc pattern recognition, thereby addressing the operational limitations of traditional FDSs. In achieving this aim, a set of objectives was laid down:

1) Choosing and adapting an HTAP benchmarking suite, including setting up both separate OLTP and OLAP components for testing.
2) Seamlessly running both operational and analytical operations on an HTAP computational set-up.
3) Evaluating the performance of various runs using the chosen HTAP configuration and different data movement scenarios.

III. LITERATURE REVIEW

A. Fraud Detection in Banking

Fraud is defined as any activity that relies on deception to achieve a gain. Banks are among the most obvious targets for fraudsters seeking financial gain. Given the complexity of modern fraud schemes and the sophistication of fraudulent behaviour, the financial industry faces some of the toughest detection and prevention challenges, necessitating advanced systems that monitor multiple sources of data.

A recent study revealed that fraud in the banking industry has become a matter of grave concern for almost all countries across the globe, causing significant financial and non-financial damages to banks, customers, other stakeholders and economies [2]. Fraud in banking also results in reputational damage and compliance challenges. Regulatory compliance further necessitates robust Fraud Detection frameworks to protect sensitive data and ensure adherence to legal standards and frameworks.

B. Fraud Detection Systems Development

Traditional FDS primarily rely on rule-based approaches, which flag suspicious activities based on predefined criteria, such as unusually high transaction amounts, frequent transactions, or geographical discrepancies [3]. These static rules struggle to adapt to the continuously evolving tactics employed by fraudsters.

Machine Learning (ML) techniques are increasingly being integrated into traditional FDS to address some of these limitations, as discussed by Minastireanu [4]. These models typically employ supervised learning models trained on historical data [5]. They exhibit reduced efficacy against novel fraud schemes that deviate significantly from historical fraudulent patterns. Additionally, these ML approaches often require substantial time and resources for periodic retraining and validation to maintain their relevance.

A further notable limitation of traditional systems is the separation of the transactional data, which is used in day-to-day activities, from the analytical data utilised in reporting and analysis. These distinct operational environments are often optimised differently, as discussed by Camilleri et al. [16], impeding the integration of real-time transaction analysis with historical data evaluation. Lastly, the broader advancement of Fraud Detection methodologies suffers due to restricted knowledge-sharing within the public domain.

C. Database Management Systems for Fraud Detection

DBMSs are critical components in Fraud Detection architectures, serving as the backbone for processing large volumes of transactional and analytical data. As banks grapple with increasing transaction volumes and sophisticated fraud schemes, traditional DBMSs face limitations, prompting the exploration of more advanced DBMS architectures that can provide near-real-time responses, robust scalability, and comprehensive data integration.

A DBMS ensures that data remains available to users and applications, handles increasing volumes of data without loss of data consistency, and can scale to meet rising demand. DBMSs play an increasingly crucial role in supporting Fraud Detection mechanisms in banking. The ability of a DBMS to process, store, and retrieve data efficiently in a structured manner can significantly improve the quality of Fraud Detection efforts.

At its core, an FDS must rapidly ingest large volumes of heterogeneous data, enforce stringent data quality and integrity constraints, and support ad hoc analytical queries. Nevertheless, relational DBMSs have encountered bottlenecks in performance and scalability, especially in transactional processing [6]. Detecting patterns of suspicious transactions requires the ability to query and analyse historical data alongside real-time transactions.

To address these challenges, the DBMS landscape has diversified into multiple directions. Noticeably, columnar with in-memory analytical platforms are enabling HTAP by integrating OLTP and OLAP workloads within a single engine. By maintaining online materialised views or employing Multi-Version Concurrency Control (MVCC) optimised for mixed workloads, HTAP systems permit on-the-fly computation of fraud scores without the latency penalties of ETL pipelines [7].

Practical performance depends heavily on real-world data patterns, which are driven not just by the underlying schema

but by unpredictable end-user behaviour, which make static optimisation insufficient. This unpredictability creates additional challenges in scaling transactional throughput while preserving data consistency guarantees.

D. Database Architecture for Fraud Detection

In many architectures, OLTP and OLAP systems operate and are set-up in isolation. OLTP databases are optimised for light transactions with few tuples in its scope, handling inserts, updates and deletes with ACID (Atomicity, Consistency, Isolation and Duration) guarantees, whereas OLAP engines are tuned for complex, read-heavy queries against large datasets. To bridge these worlds, ETL pipelines are constructed to periodically move data from OLTP systems into purpose-built Data Warehouses (DWH) to support OLAP. This also introduces latency: data freshness is bounded by the ETL cadence, and the overhead of maintaining dual schemas with the pipeline code can be substantial [8].

There are two major challenges organisations face to maintain a data infrastructure that caters for their data driven decision making, namely; (i) performance in terms of query response time, transactional throughput, and resilience to computational failures, and (ii) moving data from internal and external data sources to build an integrated, synchronised, business-process focused, and time-variant data repository, i.e., a Data Warehouse [9].

If data freshness becomes a requirement, as in a banking FDS, then the architecture with ETL does not meet it. HTAP architectures seek to collapse the boundary between OLTP and OLAP by introducing a computational set-up that supports both workloads within a single, integrated engine and instance.

This unified approach can dramatically reduce data movement overheads, eliminate ETL-induced data staleness, and simplify system maintenance. However, it does demand advanced engine optimisations and comes at a loss of the decoupling and staging flexibility offered by ETL pipelines. DWHs and OLAP asynchronous updates are typically not suitable for real-time Fraud Detection, as they are designed to process large datasets in batches, resulting in delays that could allow fraudsters to act before fraud is detected, as OLAP is querying stale (i.e., not fresh) data [10].

E. HTAP Architectures and Techniques

HTAP, attributed to Gartner in 2014 by Zhang et al. [8], describes a database architecture that unifies transaction processing and analytics in near real-time, allowing OLTP and OLAP workloads to run side by side without undue interference, as seen in Figure 1. By collapsing separate systems and eliminating complex ETL pipelines [11], HTAP simplifies data management. However, supporting both transactional and analytical demands simultaneously remains challenging, given their inherently different performance and resource requirements.

Overall, supporting both transactional and analytical queries on the same dataset increases the risk of data contention in HTAP systems, necessitates careful management of isolation levels, concurrency controls, and workload balancing strategies is essential.

HTAP remains a conceptual model rather than a formally standardised architecture, and there is currently no broad consensus on the best way to implement it. One common architecture is the primary row store with an in-memory column store, where the primary storage is a row-oriented database optimised for transactional operations, and an in-memory column store used to handle analytical queries. This design facilitates real-time analytics without compromising transactional performance.

The choice of architecture depends on specific application requirements, workload characteristics and resources available. For instance, high-volume transactional systems favour OLTP for fast, consistent writes, while analytics-heavy workloads benefit from OLAP's read-optimized design. Hybrid systems like HTAP are more suitable when near real-time insights are needed without sacrificing transactional integrity, such as in fraud detection scenarios. Additionally, resource constraints (e.g., limited memory or compute power) may dictate the use of simpler architectures, whereas larger organizations with more capacity can adopt in-memory or distributed architectures for greater responsiveness and scale.

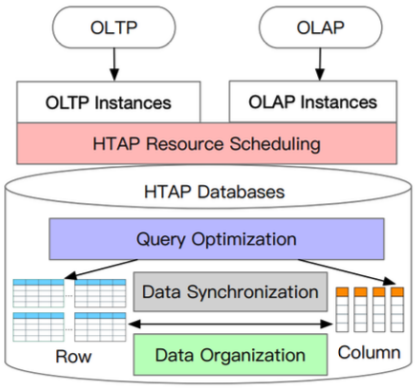

Figure 1. HTAP System Overview [15].

The primary challenge in an HTAP system is to facilitate the smooth flow of data from all these different sources into a unified platform where both transactional and analytical operations can occur. Ensuring that transactional updates and analytical views remain tightly consistent is critical for applications that demand ACID guarantees, such as Fraud Detection in banking. Thus, HTAP platforms in banking must implement strong consistency mechanisms to synchronise updates from OLTP to OLAP stores with minimal latency and without sacrificing throughput.

F. Benchmarking

Benchmarking database systems underpins objective performance assessment, offering repeatable tests that measure metrics, such as throughput, latency or completion time against standardised workloads [11]. A typical benchmark defines a data schema, data volume, workloads (queries and transactions) and an evaluation criterion. These tests serve multiple purposes: identifying raw horsepower

through intensive tuning; comparing enhancements or configuration changes; and contrasting architectures or hardware choices under controlled conditions. Fairness, consistency and conservatism in benchmark design are essential to yield meaningful insights.

Several HTAP-specific benchmarks have emerged [13]. Early efforts, such as CH-benCHmark extend the Transaction Processing Performance Council's (TPC) benchmarks TPC-C and TPC-H by executing transactional and analytical workloads on the same dataset, thereby revealing workload contention effects [14]. However, these benchmarks still lack the depth required for domains like banking, where anomaly detection and time-sensitive analytics are paramount.

HyBench [15] was developed to address these precise challenges. It integrates realistic banking transactions alongside analytical tasks like trend analysis. Its workload comprises 18 transactional operations and 13 analytical queries executed under mixed scenarios, mirroring real-world HTAP use cases more closely than its predecessors. By focusing on financial workloads and anomaly-driven scenarios, HyBench offers a comprehensive evaluation framework that blends transactional throughput, analytical query performance and data freshness. Its extensible design allows practitioners to adjust parameters, incorporate new metrics or inject domain-specific logic, making it a versatile tool for advancing HTAP research and guiding production deployments in latency-sensitive environments.

IV. DESIGN, IMPLEMENTATION AND TESTING

G. System Design

HyBench forms the foundation of our evaluation framework, simulating both transactional and analytical workloads on a unified schema mimicking typical banking entity, such as customers, accounts, transactions, loans and companies, with SFs of 1x, 10x and 100x factors to approximate 1 GB, 10 GB and 100 GB of base data, respectively. A harness orchestrates the workflow: it restores a clean database snapshot, vacuums and analyses tables to eliminate fragmentation, pre-warms the cache using a representative query, then launches concurrent clients for:

- Transactional Processing (TP) workloads, with high-volume, short-lived OLTP operations
- Analytical Processing (AP) workloads, with long-running, scan-intensive OLAP-style queries
- Hybrid Processing (XP) mixed workloads, combining both TP and AP workloads

All tests were conducted on a standalone machine (macOS, M3 Pro, 16 GB RAM, 500 GB SSD), using PostgreSQL 16, with pg_stat_statements enabled to record detailed query metrics. The Python 3.13.2 harness performed orchestration and monitoring, while Java components (via OpenJDK 21) drove the HyBench workload. Key PostgreSQL configurations were tuned, allowing up to 100 concurrent connections, allocating 4 GB each to shared_buffers and effective_cache_size, disabling autovacuum, and setting work_mem to 64 MB. These configurations were chosen based on the available hardware capacity and industry best practices.

During each benchmark, we gathered metrics from both HyBench logs and the Python harness, e.g., query performance data—including total and mean execution times (ms), execution counts, rows returned, shared block hits and reads, and the query text—table activity counts for inserts, updates and deletes, lock information detailing lock types and their frequencies, transactional throughput figures for committed transactions per table, as well as total transactions (committed plus rolled back), and storage statistics listing schema and table names alongside total table and index sizes.

Together, these KPIs offer a comprehensive view on resource use and contention under concurrent OLTP/OLAP conditions. To unify these dimensions, HyBench's creators introduced the H-Score. This unified metric incorporates Transactions Per Second (TPS), Queries Per Second (QPS), mixed workload throughput (XPS = TPS + QPS for mixed workloads), data freshness (f_s) and Scale Factor (SF) to yield an overall performance rating. The H-Score is defined as the geometric mean of all throughputs, multiplied by the SF and divided by the freshness metric, as in (1).

$$H\ Score\ =\ SF\ \times \frac{\sqrt[3]{TPS \times QPS \times XPS}}{f_s+1} \qquad (1)$$

H-Score is beneficial as solely relying on one aspect cannot reflect the true HTAP performance. The five components in H-Score are widely recognised by benchmarking suites as the most important factors for quantifying the HTAP performance [10].

H. Implementation

We defined 11 core parameter dimensions to explore various workload mixes and data volumes, producing 144 unique configurations, as seen in Table 1, but we ultimately retained only 96 combinations (i.e., the 1x and 10x SF ones) after observing inconsistent KPI behaviour at 100x, namely scalability limits in HyBench's threading and PostgreSQL's I/O performance on given computer set-up.

By adjusting these parameters, we could evaluate scenarios with standard transactional-heavy loads (3:1 TP:AP) and more analytics-intensive mixes (3:2 TP:AP), as well as stress-test at 100x SF.

A Python harness was built to minimize manual intervention. The script handled the below process:

1) Generate the HyBench '.props' files programmatically for every parameter combination.
2) Restore and reset the database from a known backup via 'pg_restore' utility, then execute VACUUM FULL ANALYZE on all tables to rebuild storage and refresh Data Dictionary (DD) statistics.
3) Prewarm the cache using a realistic query that joins key tables and exercises index and sequential scans in parallel, ensuring relevant pages are memory-resident.
4) Launch HyBench with a '.props' configuration file, while a background process queries the DD views, pg_stat_statements, pg_locks, pg_stat_user_tables and pg_stat_database every 60s for live metrics.
5) Parse the logs and plot key performance indicators.

TABLE I. HYBENCH PARAMETER CHOICES

Parameter	Set Value	Description
sf	1x, 10x, 100x	Scale factors for the table data
at_percentages	(35,25,15,15,7,3), (3,7,15,15,25,35), (10,10,20,20,20,20)	AT Ratio (sum = 100%)
apclient	10	AP concurrency
tpclient	15, 30	TP concurrency
fresh_interval	150	Freshness evaluation is done every (xpRunMins/150) s
apRunMins	5, 10	AP evaluation time
tpRunMins	5, 10	TP evaluation time
xpRunMins	5, 10	XP evaluation time
xapclient	10	XP-ATS concurrency
xtpclient	15, 30	XP-IQS concurrency
distribution	Uniform	Data distribution at generation phase

I. *Testing*

We began with unit tests of individual harness modules (e.g., DB housekeeping, monitoring). Next, integration tests ran end-to-end workflows using varied '.props' files to validate the sequence: restore → housekeeping → prewarm → workload → monitoring. Logs from PostgreSQL and HyBench were cross verified to ensure consistency, accounting for differences in granularity (e.g., DBMS internal vs. atomic operations launched by HyBench). All runs were performed in a controlled environment with server resource prioritisation to avoid external interference. Three main scenarios were executed – each taking 24 hours:

1) 1x: Establish expected performance and tune PostgreSQL parameters to avoid configuration-induced artefacts.
2) 10x: Stress test concurrency, observe degradation in freshness and throughput, and validate that HTAP sustains desired freshness under mixed workloads.
3) 100x: Identify limits of the harness and the hardware set-up available for the DBMS, revealing thread-management issues, Java Database Connectivity termination overheads, and I/O saturation that rendered metrics unreliable (consequently it was discarded for evaluation).

For each run, we generated data (via HyBench's gendata module), created indexes and tables, and ran the full benchmark cycles. Post-run validation included checking table and index sizes, transaction counts and comparing relative throughput trends against published HyBench with PostgreSQL baselines to confirm that our findings aligned qualitatively with prior results.

V. EVALUATION

Across both 1x and 10x data volumes, the HTAP setup consistently delivered sub-20 ms data freshness, demonstrating its ability to service the latest transactional changes to analytical queries almost instantaneously. At 1x scale, the F-Score typically ranged from 2 ms to 12 ms, with occasional peaks near 12 ms under high contention; at 10x, peaks rose only slightly—up to around 15 ms—confirming robust freshness even under heavier mixed loads. This encouraging performance underpins real-time use cases where even small delays can blindside FDS.

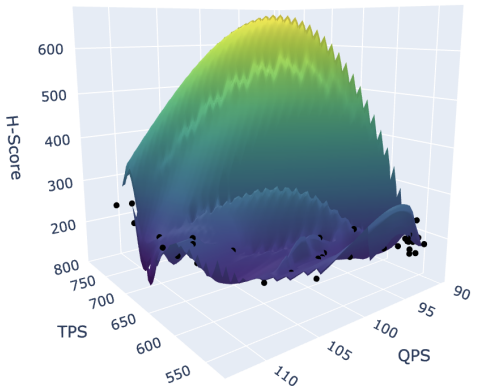

Figure 2. TPS vs QPS vs H-Score over all 1x Runs.

The composite H-Score further highlights the trade-offs inherent in unified HTAP processing. At 1x, H-Scores were higher but volatile, spanning 180 to 260, reflecting bursts of transactional and analytical contention (see the sharp ridges and valleys in Figure 2). Conversely, 10x runs yielded lower but much more stable H-Scores (about 18–28), indicating that increased data volume smooths performance variability through more effective MVCC snapshot isolation and adaptive resource scheduling (as shown by the more compressed surface in Figure 3).

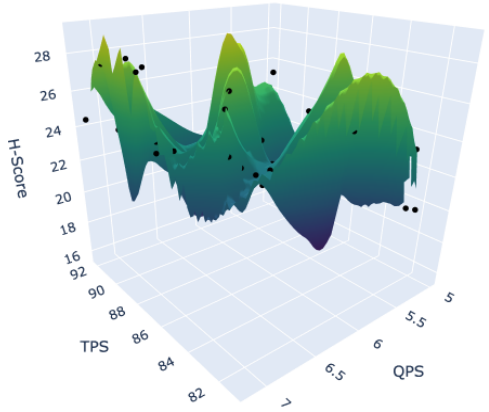

Figure 3. TPS vs QPS vs H-Score over all 10x Runs.

Although running OLTP and OLAP workloads independently on the same system can yield higher raw TPS or QPS in isolation, this approach still lacks the integration needed for real-time analytics and introduces delays when used in sequence. In contrast, the HTAP unified architecture eliminates these delays by supporting concurrent transactional and analytical workloads within a single engine.

Mixed-mode throughput in our HTAP configuration was approximately 3.5x lower than the sum of isolated workloads at SF 1x, and around 4x lower at SF 10x, due to shared resource contention under mixed loads. However, these throughput trade-offs are more than offset by the removal of data staleness, pipeline complexity, and maintenance overhead, resulting in a streamlined, low-latency platform ideally suited to modern, mission-critical analytics.

VI. CONCLUSION

The overall objective of our evaluation was to investigate the feasibility of employing an HTAP architecture capable of handling mixed workloads, while also delivering near real-time data insights within a fraud detection scenario. The technique to evaluate this feasibility was by using an HTAP benchmarking suite, while simultaneously building a harness in order to be able to orchestrate the entire process, keeping reproducibility and fairness across all runs.

Our runs show that, with proper configuration, HTAP systems can consistently uphold demanding Service Level Agreements by keeping sub-second data freshness while handling heavy mixed workloads at scale. The benchmarking methodology offers a practical guide for rolling out HTAP in live environments. However, one must take into consideration contention between long-running analytical queries and high-frequency transactions which led to latency spikes–highlighting the need for fine-grained concurrency control and adaptive resource scheduling. Early attempts using default database configurations significantly underperformed, reinforcing the necessity of tailored tuning and proactive monitoring. To capitalise on these advantages, data architects should start incorporating HTAP-aware access patterns and concurrency controls starting from the application design phase.

Future research could expand in several directions. Pushing tests to HyBench's higher scale factors would shed light on I/O behaviour, buffer management and concurrency under extreme loads and therefore better sizing of the instance. Exploring in-memory databases with persistent Non-Volatile RAM logging, as well as evaluating DBMS auto-tuning features for adaptive query optimisation, could further improve low-latency analytics. Standardising HyBench and adding support for varied data distributions, refining freshness metrics, and harmonising threading models is a critical need. Finally, extending the harness presented here for domain-specific microbenchmarks would create a unified framework for HTAP evaluation across sectors.

Looking ahead, the full adoption of HTAP architectures holds significant promise, but not without challenges. On the one hand, HTAP offers a path to simpler architectures, fresher data, and faster insights extraction, aligning closely with modern regulatory, operational, and customer expectations. On the other hand, widespread adoption will require rethinking application patterns, retraining engineering teams, and overcoming vendor lock-in as HTAP maturity varies across platforms. Moreover, while HTAP simplifies data pipelines, it shifts complexity into query optimisation, workload isolation, and configuration management; domains that still require advanced expertise and careful management. If these challenges can be addressed, HTAP could become a cornerstone for near real-time, data-driven decision-making in finance and beyond.

REFERENCES

[1] O. O. Elumilade, I. A. Ogundeji, G. O. Achumie, and H. E. Omokhoa, "Enhancing Fraud Detection and forensic auditing through data-driven techniques for financial integrity and security," Journal of Advanced Education and Sciences, vol. 1, no. 2, pp. 55–63, 2021.

[2] D. Mangala and L. Soni, "A systematic literature review on frauds in banking sector," Journal of Financial Crime, vol. 30, no. 1, pp. 285–301, 2023.

[3] N. Faisal, J. Nahar, N. Sultana, and A. A. Mintoo, "Fraud Detection in Banking Leveraging AI to Identify and Prevent Fraudulent Activities in Real-Time," Journal of Machine Learning, Data Engineering and Data Science, vol. 1, no. 01, pp. 181–197, 2024.

[4] E. A. Minastireanu and G. Mesnita, "An Analysis of the Most Used Machine Learning Algorithms for Online Fraud Detection.," Informatica Economica, vol. 23, no. 1, 2019.

[5] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A Survey," Journal of Network and Computer Applications, vol. 68, pp. 90–113, 2016.

[6] S. A. Ionescu, V. Diaconita, and A. O. Radu, "Engineering Sustainable Data Architectures for Modern Financial Institutions," Electronics, vol. 14, no. 8, p. 1650, 2025.

[7] Z. Zhang, A. Megargel, and L. Jiang, "Performance Evaluation of NewSQL Databases in a Distributed Architecture," IEEE Access, 2025.

[8] C. Zhang, G. Li, J. Zhang, X. Zhang, and J. Feng, "HTAP Databases: A Survey," IEEE Transactions on Knowledge and Data Engineering, 2024.

[9] W. Lehner, "Merging OLTP and OLAP–Back to the Future: (Panel)," in International Workshop on Business Intelligence for the Real-Time Enterprise, Springer, 2009, pp. 171–173.

[10] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker, The data warehouse lifecycle toolkit. J Wiley, 2008.

[11] D. Beyer, S. Löwe, and P. Wendler, "Reliable benchmarking: Requirements and solutions," Int. Journal on Software Tools for Technology Transfer, vol. 21, no. 1, pp. 1–29, 2019.

[12] TPC-C Homepage — tpc.org, https://www.tpc.ord/tpcc/. [Retrieved: June, 2025]

[13] S. Leutenegger "A modeling study of the TPC-C benchmark," ACM Sigmod Record, vol. 22, no. 2, pp. 22–31, 1993.

[14] R. Cole et al., "The mixed workload CH-benCHmark," in Proceedings of the Fourth International Workshop on Testing Database Systems, 2011, pp. 1–6.

[15] C. Zhang, G. Li, and T. Lv, "HyBench: A New Benchmark for HTAP Databases," Proceedings of the VLDB Endowment, vol. 17, no. 5, pp. 939–951, 2024.

[16] C. Camilleri, C. Vella and V. Nezval, "HTAP with Reactive Streaming ETL," JCIT, vol. 23, no. 4, pp. 1-9, 2021.