ISSN 2085-4552

Trends and Keyword Networks in Machine Learning-Based Click Fraud Detection Research

Kevin[1], Aditiya Hermawan[2]
[1,2] Dept. of Informatics Engineering, Buddhi Dharma University, Tangerang, Indonesia
[1]kevinawi1112@gmail.com, [2]aditiya.hermawan@ubd.ac.id

Accepted 25 June 2025
Approved 30 June 2025

***Abstract*— The rapid advancement of the digital economy has significantly increased the use of online advertising while concurrently giving rise to critical challenges, particularly in the form of click fraud—a manipulative act that harms advertisers by generating fraudulent clicks on digital advertisements. As click fraud attack patterns grow increasingly complex, machine learning (ML)-based research has emerged as a principal approach for detecting and mitigating these threats. This study aims to map the research landscape of ML-based click fraud detection through a bibliometric analysis to identify publication trends, patterns of international and institutional collaboration, and key thematic domains within this field. Employing a bibliometric methodology, the study analyzed 61 publications retrieved from Dimensions.ai spanning the years 2015–2024. The data were collected, refined using OpenRefine, and visualized with VOSviewer to examine keyword co-occurrences and research trends. The findings reveal a marked increase in publication volume since 2019, with dominant contributions from India, China, Saudi Arabia, and the United States. Furthermore, four principal research clusters were identified: cybersecurity, the relationship between click fraud and the digital advertising industry, dataset processing and evaluation techniques, and the development of ML-based detection systems. Each cluster offers practical contributions in areas such as system protection strategies, ad budget optimization, improved detection accuracy, and the development of scalable, real-time detection solutions. Recent trends highlight growing scholarly interest in model performance evaluation and the challenges posed by class imbalance (class skewness). This study concludes that more effective data management and the development of adaptive ML models capable of addressing evolving attack patterns are pivotal for future research. By providing a clearer mapping of current trends, this study aims to support the scientific community in developing more accurate and efficient click fraud detection strategies, thereby strengthening the integrity of the global digital advertising ecosystem.**

***Index Terms*— Click Fraud; Machine Learning; Bibliometric Analysis; Fraud Detection; Digital Advertising**.

I. INTRODUCTION

Over the past few decades, the digital economy has grown rapidly worldwide [1]. This sector has also become a key transmission hub in the economic system, contributing significantly to global economic growth [2]. One of the main drivers of this growth is technological advancement [3]. The development of technology has enabled companies to reach consumers more effectively through digital platforms.

However, alongside this rapid growth, new challenges related to digital security have emerged, particularly in the form of Click Fraud, a manipulative act that generates fraudulent clicks on advertisements with the intent of harming advertisers [4]. A 2020 study by the University of Baltimore found that click fraud caused losses exceeding $35 billion [5]. Click fraud not only results in substantial financial losses for advertisers but also undermines the integrity of the digital advertising ecosystem as a whole. To address this threat, technology-based solutions are required. Detecting click fraud generally relies on machine learning models, which have become one of the most effective approaches due to their ability to learn complex behavioral patterns and identify subtle anomalies that signal fraudulent activity [6].

The application of ML techniques in detecting click fraud has received significant attention in recent years. ML provides the capability to analyze complex data patterns and identify anomalies that traditional methods might overlook. A study by Aljabri investigated the application of machine learning models to distinguish between human and bot click behaviors in pay-per-click (PPC) advertising. The results showed that while all models achieved strong performance, the Random Forest algorithm consistently outperformed others across all evaluation metrics, indicating its robustness in detecting fraudulent ad-click activity [7]. Additional research also highlighted that ensemble methods can further enhance detection performance [8] .

Despite the increasing number of studies applying machine learning for click fraud detection [9], there is a lack of systematic synthesis regarding how research in this field has evolved, what methods are predominantly used, and which conceptual domains remain underexplored. Previous reviews have tended to focus on algorithmic performance or case-specific implementations, rather than providing a macro-level mapping of the intellectual structure of the field. In contrast, bibliometric studies in other fraud detection domains, such as financial fraud or healthcare fraud, have provided broader overviews of research trends. These studies often focus on general approaches or dominant techniques, such as decision trees, SVMs, or neural networks, but fail to provide a detailed mapping of publication trends or global research collaboration patterns. Therefore, this study fills an important gap by conducting a bibliometric analysis to uncover research trends, influential themes, and methodological patterns in the intersection of machine learning and click fraud detection. The findings are expected to inform both academic research agendas and practical implementations in digital advertising security.

This study aims to address the identified research gap by conducting a bibliometric analysis of scholarly literature focused on click fraud detection using machine learning (ML) techniques. Bibliometric analysis is a widely adopted approach with high methodological validity for examining large bodies of academic literature. This method enables researchers to trace the historical development of a scientific discipline and to identify emerging directions and novel themes within the field [10]. Unlike general fraud detection bibliometrics, which primarily examine techniques applied to broader domains like finance or healthcare, this study focuses on the specific context of click fraud, offering in-depth insights into research trends unique to the digital advertising ecosystem. By mapping the evolution of research in this domain, the study seeks to provide in-depth insights into the current and prospective trajectories of scholarship in click fraud detection. Furthermore, a better understanding of prevailing trends and research patterns may contribute to the development of more effective strategies for detecting and preventing click fraud, thereby reinforcing the integrity of the digital advertising ecosystem.

To achieve these objectives, this study seeks to answer the following research questions:

1. How have publication trends in Click Fraud detection using Machine Learning evolved over time, both in terms of the number of publications and collaboration patterns among countries and institutions?
2. What are the key research topics and keyword co-occurrence patterns in ML-based Click Fraud detection studies, as identified through bibliometric analysis?

How has the research focus on ML-based Click Fraud detection shifted over time, particularly in terms of keyword relationships and emerging topics in recent years?

II. METHOD

The procedure used to conduct this research consists of five stages. These stages are as follows: Data Collection, Data Cleaning, Data Visualization, Data Analysis, and Report Writing. Fig 1 illustrates how this procedure should be carried out in more detail.

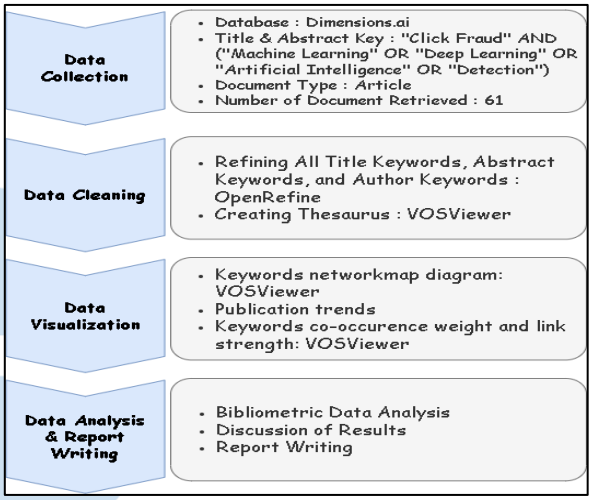

Fig 1. Research methodology

A. Data Collection

The data was obtained from the Dimensions.ai website as part of the data collection phase in the form of a CSV file, the selection of Dimensions.ai as the primary data source was based on its open and freely accessible nature, which facilitates the independent execution of bibliometric analysis. The applied publication year restriction spans from 2015 to 2024, covering a 10-year range. The search query used was: TITLE-ABSTRACT ("click fraud" OR "ad fraud") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "detection"). This search was restricted to journal articles only. Using this search technique, a total of 61 articles were retrieved. Although relatively limited in quantity, the 61 publications included in this study were manually screened and curated to ensure high thematic relevance to the specific domain of click fraud detection using machine learning. Broader search queries using general terms such as 'fraud' produced a considerable number of irrelevant results—covering areas like financial fraud, healthcare fraud, and identity theft—which would have diluted the semantic focus of the analysis. Therefore, this study deliberately prioritized semantic precision over corpus size, a methodological trade-off commonly accepted in bibliometric analyses of niche or emerging topics. Moreover, Glänzel and Moed [11] suggest a rule-of-thumb minimum of 50 documents to ensure approximate properties such as normality in the

distribution of means and relative frequencies. With 61 highly relevant articles, this study meets that threshold and maintains sufficient statistical integrity for meaningful co-word and thematic mapping.

B. Data Cleaning

This data cleaning phase aimed to ensure more accurate exploration of bibliometric and bibliographic data, as well as to enable improved visualization and interpretation of the results [10]. All keywords used in the Title and Abstract fields were standardized using OpenRefine. OpenRefine facilitated the detection of semantically similar keywords by identifying lexical variations within the dataset, thereby supporting the standardization and consolidation of terms that are conceptually identical but expressed differently. This process had a significant impact on enhancing the accuracy and integrity of the keyword co-occurrence network structure, as the merging of redundant terms prevented the fragmentation of thematic clusters that could otherwise distort the conceptual mapping. Consequently, the resulting network visualizations more accurately reflect the dominant themes within the literature and strengthen the validity of interpretations regarding topical interconnections within the analyzed research corpus. Table I presents examples of keyword standardizations performed during the data cleaning process using OpenRefine.

TABLE I.KEYWORD STANDARDIZATION EXAMPLES

Original Keyword	Standardized Keyword
Prediction, predicting, predict, predicted	prediction
Demonstrate, demonstrated, demonstrates	demonstrate
Classifier, classfiers, classifies, classfier's	classifier
Fraudster, fraudsters, fraudster's	fraudsters

C. Data Visualization

The Data Visualization Phase was carried out by constructing a network map based on keyword co-occurrence from the analyzed articles using VOSviewer. This phase aimed to identify relationships between keywords in the dataset and explore conceptual linkages within this research field.

At this stage, the minimum keyword co-occurrence threshold was set at 6, resulting in the selection of 79 keywords from a total of 1,697 available terms. The selection of a threshold of six was not arbitrary; rather, it aligns with established bibliometric practices and is theoretically grounded in the thresholding formula introduced by Donohue [12], as operationalized in subsequent studies such as [13], [14]. This method estimates the optimal boundary for distinguishing high-frequency keywords based on the distribution of singleton terms within the corpus. By applying this threshold, the present study adheres to a well-documented standard in co-word analysis, which ensures analytical consistency and avoids distortions caused by low-frequency noise.

To confirm its appropriateness, a limited sensitivity trial was conducted by comparing alternative thresholds. When the threshold was reduced to four, 97 keywords were retained—exceeding the recommended upper boundary of 67 high-frequency terms as per the Donohue model, and introducing considerable lexical noise. Conversely, raising the threshold to eight and ten produced only 57 and 6 keywords, respectively—both falling below the recommended inclusion range and omitting key conceptual terms. Consequently, additional sensitivity testing was deemed unnecessary, as the threshold has been validated and widely adopted in comparable bibliometric investigations.

Furthermore, out of the 79 identified keywords, only 60% (47 keywords) were used as the final threshold. In bibliometric analysis, the 60% threshold is the default setting in VOSviewer and is considered a best practice [15].

Additionally, Python with the Plotly library was used to generate bibliometric data visualizations, such as the total publication distribution by country, which helps identify the most productive countries in this research domain. Moreover, publication trends over the years were analyzed to examine research developments within a specific time frame.

D. Data Analysis and Report Writing

The final phase of this study consists of data analysis and report writing. The bibliometric data presented in the data visualization phase is then evaluated and interpreted based on the articles included in this study. The interpretation of results is based on the bibliometric data visualizations generated in the previous phase, including the analysis of the network map diagram, which was constructed using the co-occurrence of article keywords. The findings, discussion, and conclusions of this research are then summarized in a comprehensive report, ensuring a clear understanding of the trends and conceptual linkages identified in the study.

III. Result And Discussion

This section presents the results of the bibliometric analysis on Click Fraud Detection research. The analysis was conducted to identify publication trends over time, publication distribution by country, and the most frequently used machine learning methods for click fraud detection. These findings provide a comprehensive overview of the research developments in this field, including the number of publications, citation impact, and the dominant techniques in current scientific approaches.

A. Publication Trends Over Time

The publication trend analysis indicates that research on Click Fraud Detection has experienced a significant increase after 2019. As shown in Fig. 2, the

highest number of publications was recorded in 2022, with 14 articles published that year. Specifically, the number of publications grew by 55.56% from 2020 (9 articles) to 2021 (14 articles), reflecting a sharp surge in interest in this area. The trend reveals a steady increase, with an annual growth rate of approximately 30% from 2019 to 2022, followed by a slight decline in 2023 and 2024. This growth indicates the expanding importance of click fraud detection in the context of the digital economy. Despite an average annual growth rate of +14.9%, the trend is heavily skewed by extreme outliers, indicating that the field's growth is non-linear and highly volatile. This pattern may reflect a combination of dataset limitations (e.g., incomplete indexing for 2024), external disruptions (such as funding shifts or academic redirection), and possible saturation in the core area of click fraud detection.

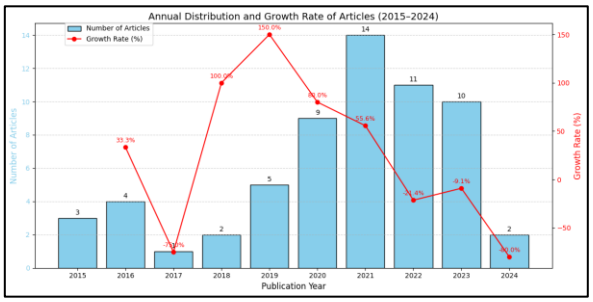

Fig 2. Annual distribution of articles on click fraud detection (2015–2024)

The period from 2018 to 2019 saw research in this area still in its early exploratory phase, with a relatively low number of publications. However, a sharp increase in publications occurred from 2020 to 2022, marked by significant growth in research output, a rise in citation impact, and stronger interconnections among studies in this domain.

B. Geographic Trends of Publications

The analysis results indicate that publications on this topic are globally distributed, with certain countries contributing more significantly than others. Fig. 3 presents the top eight countries with the highest number of publications in click fraud detection research, along with the exact number of publications per country.

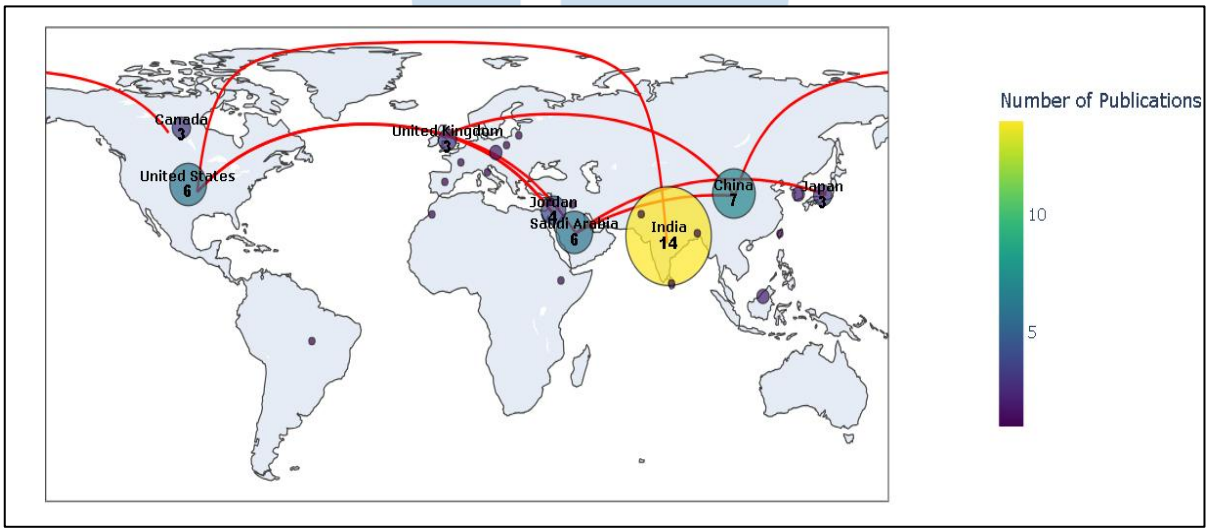

Fig 3. Top eight countries with the highest number of publications in click fraud detection research

Fig. 3. Top eight countries with the highest number of publications in click fraud detection research. The number of publications is indicated for each country: India (14), China (7), Saudi Arabia (6), the United States (6), and others including Jordan, and several European nations.

India emerges as the leading country in this research domain, contributing a total of 14 publications, which accounts for 23% of the total publications in this domain. India's dominance in this field can be attributed to its rapidly growing information technology industry. China, contributing 7 publications (approximately 11% of the total), and Saudi Arabia with 6 publications (about 10%), also demonstrate significant interest in click fraud detection, particularly in the context of protecting their digital advertising ecosystems. The United States, as a global hub for digital technology and advertising, has contributed 6 publications (around 10%), indicating continued academic and industrial engagement in this research area.

Other countries, such as Jordan and several European nations, have also made notable contributions to this field, though their contributions are smaller in scale. The overall distribution of publications highlights a growing global interest in ML-based click fraud

detection, with nations not only from large digital advertising markets but also those prioritizing cybersecurity and the efficiency of digital advertising systems.

C. Network Map Diagram Analysis

The network map diagram based on keyword co-occurrence in the analyzed articles was generated using VOSviewer. The term "keyword co-occurrence" refers to the frequency with which a keyword appears across multiple publications. The minimum occurrence threshold can vary significantly depending on the research objectives. A lower threshold results in more keywords being displayed, while a higher threshold reduces the number of displayed keywords.

Researchers extracted 1,697 keywords from a total of 61 articles. The minimum threshold for keyword co-occurrence was set at 6 times, resulting in 79 keywords meeting the minimum requirement. Fig 4 illustrates that only 60% of the total connections among the 79 keywords—equivalent to 47 keywords—were included in the final visualization.

The weight of an item determines the size of its label and circle in the network map. The larger the weight of an item, the bigger its label and circle. The color of each item is determined by the cluster to which it belongs, reflecting thematic groupings within the dataset.

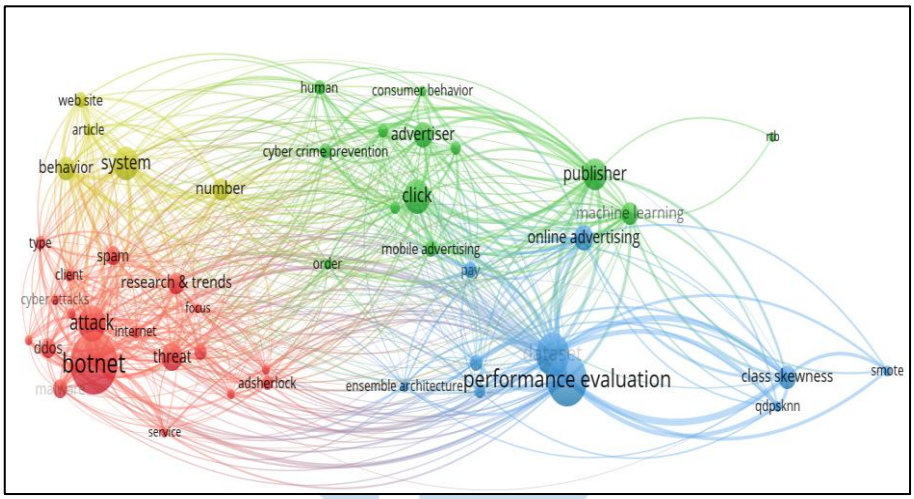

Fig 4. Network visualization of keyword co-occurrence in click fraud detection studies

Each of the four colors visible in Fig 4 represents a different cluster. The clustering approach is based on keyword co-occurrence, as detected across all analyzed articles. This indicates that elements grouped within the same cluster are more closely related to one another compared to elements outside their respective clusters. Therefore, it can be inferred that elements within the same cluster likely share a similar research focus. The details of the keywords in each cluster are summarized in Table II.

Keywords such as "botnet" and "performance evaluation" also exhibit high link strength, suggesting that detection models frequently correlate with botnet-based attack patterns and performance evaluation techniques. The presence of the term "dataset" with strong connections further indicates that data quality and dataset processing methods are key factors in the effectiveness of click fraud detection systems. Fig 5 presents the top 15 keywords with the highest co-occurrence values and total link strength.

While the co-word and cluster analyses have yielded useful thematic structures, it is important to acknowledge several limitations inherent in the bibliometric approach used. First, this study exclusively utilized the Dimensions.ai database, which although extensive in scope aggregates a wide variety of publication types and disciplines. This heterogeneity may influence the consistency and interpretive clarity of the resulting thematic patterns, particularly when compared with more curated and domain-specific bibliographic databases. To mitigate potential coverage bias and ensure methodological triangulation, future studies are encouraged to cross-validate findings using established repositories such as Scopus or Web of Science, which offer more standardized indexing criteria and peer-reviewed literature emphasis.

Second, the process of keyword normalization conducted using OpenRefine may introduce semantic ambiguity. Decisions on merging or standardizing keywords (e.g., “click fraud” vs “ad fraud”) rely partly on subjective judgment, which could influence the resulting cluster composition. Moreover, relying on author-assigned keywords may bias the analysis toward how authors frame their work, rather than capturing the conceptual content in full.

Third, while thresholding co-occurrence at six ensured analytical clarity, this choice may have excluded emerging but low-frequency terms that are thematically significant. This reflects a broader limitation of co-word analysis itself: its tendency to privilege frequency over novelty.

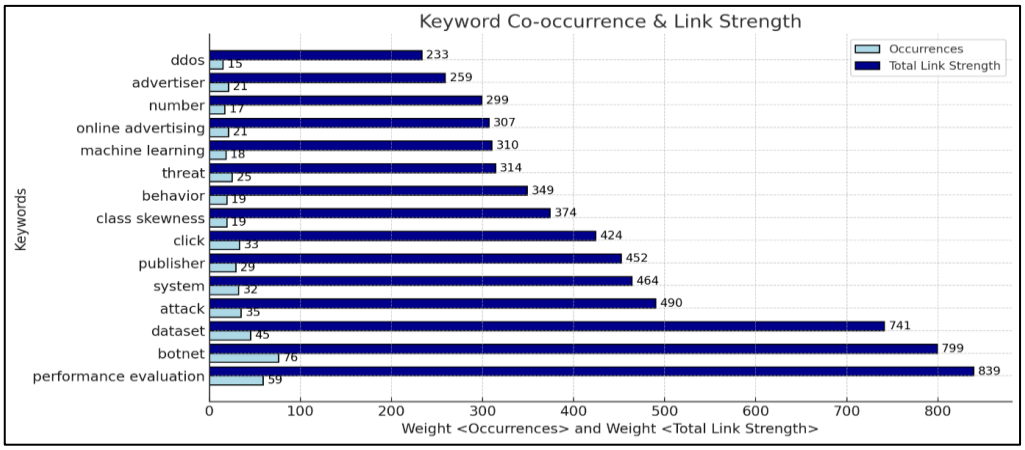

Fig 5. Top 15 keywords ranked by co-occurrence and total link strength in click fraud detection research

TABLE II. KEYWORD CLUSTERS

Cluster	Keywords	Issue
1 (19 Keywords)	adsherlock, attack, botnet, client, cyber attacks, day, ddos, focus, internet, malicious code, malware, mobile app development, online detection, phishing, research & trends, service, spam, threat, type	This cluster is highlighting the foundational concerns of network-based attack vectors and systemic vulnerabilities. This cluster aligns with real-world challenges in identifying sophisticated bot traffic and suggests the need for integrating behavioral analytics into fraud detection pipelines. Studies such as Sadeghpour and Vlajic have shown that botnets often mimic legitimate user behavior, making this cluster crucial for developing resilient detection mechanisms [16]
2 (13 Keywords)	advertiser, click, consumer behavior, cyber crime prevention, fraudulent click, human, machine learning, mobile advertising, order, parameter, publisher, revenue, rtb	This cluster comprising keywords like represents the commercial and economic dimension of the field. The prominence of these terms underscores the growing concern from advertisers and platforms over financial losses. This cluster suggests a need for models that not only detect fraud but also estimate its economic impact [17]
3 (10 Keywords)	class skewness, classification, dataset, ensemble architecture, online advertising, pay, performance evaluation, qdpsknn, smote, state	This cluster revolves around technical modeling issues, with keywords such as class imbalance, SMOTE, and ensemble learning. These terms underscore the methodological challenges in handling skewed datasets—an inherent characteristic of fraud detection tasks, where legitimate instances significantly outnumber fraudulent ones. The prominence of these terms highlights the growing emphasis on developing resilient models capable of maintaining predictive performance under such imbalance. Notably, G.S. T. et al. [18] demonstrated the effectiveness of ensemble-based methods in addressing class imbalance by leveraging the combined strengths of multiple classifiers, thereby supporting the broader adoption of hybrid learning architectures in this domain. This cluster, therefore, opens further avenues for research into meta-learning strategies and cost-sensitive algorithms tailored for rare-event prediction in click fraud detection
4 (5 Keywords)	article, behavior, number, system, web site	Cluster 4 includes terms such as model architecture, detection system, and anomaly detection, reflecting a focus on the system-level implementation of click fraud detection frameworks. This cluster serves as a bridge between theoretical algorithm development and practical engineering deployment, emphasizing the importance of scalable and explainable AI models capable of operating in real-time environments. Notably, a study by Neeraja et al. supports this direction by demonstrating that real-time ad-click fraud can be effectively identified using elementary classifiers [19]

D. Overlay Visualization

An overlay visualization was created to identify the latest research topics, as illustrated in Fig 6. The color gradient from dark to light represents the publication year, ranging from the earliest to the most recent studies. Darker blue shades indicate older research topics, while yellow shades highlight more recent discussions.

Keywords appearing in the yellow spectrum, such as "performance evaluation" and "dataset", suggest that these topics have gained increased attention in recent studies. This indicates a shift in research focus towards

evaluating the performance of click fraud detection models, emphasizing how dataset quality and characteristics influence model performance.

Additionally, the presence of the keyword "class skewness" reinforces that class imbalance in datasets has become a critical issue with direct implications for detection performance. In the early stages of research, the primary concern was to develop models capable of distinguishing between fraudulent and legitimate clicks, often evaluated using balanced or synthetically constructed datasets. However, as the field has matured, scholars have increasingly acknowledged that imbalanced class distributions are the rule rather than the exception in real-world advertising environments. This recognition has led the research community to treat class skewness not as a peripheral modeling concern, but as a core methodological and operational challenge, particularly in the context of rare-event classification and cost-sensitive decision-making.

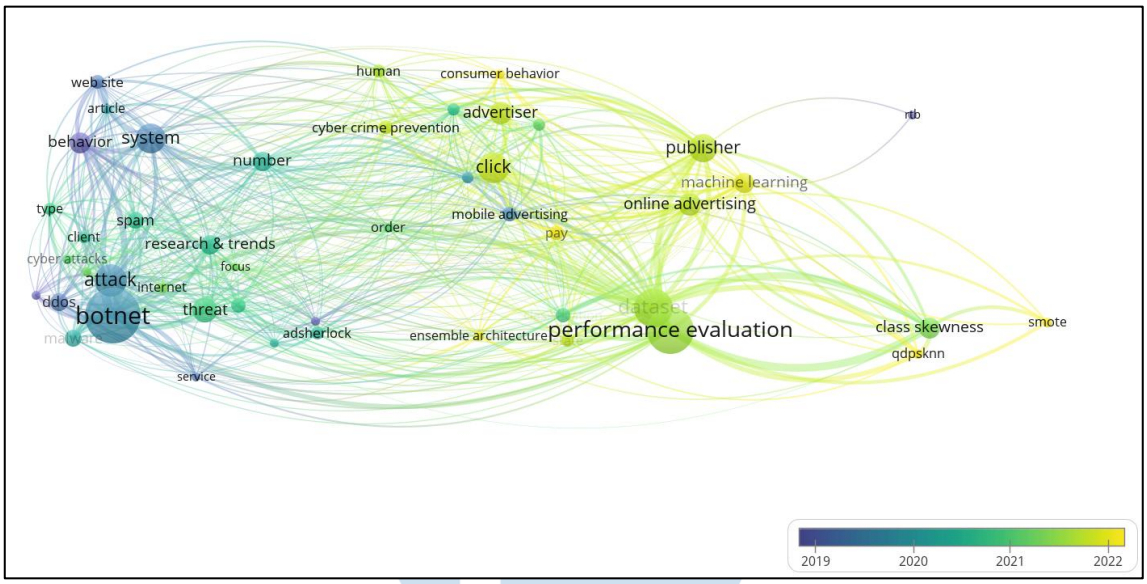

Fig 6. Overlay visualization of keyword co-occurrence in click fraud detection research

The bibliometric analysis conducted in this study reveals that research on Click Fraud Detection has experienced significant growth since 2019. This publication trend aligns with the increasing demand within the digital industry to address fraudulent activities in online advertising. Prior to 2019, the number of publications in this field remained relatively low, indicating that the research was still in an early exploratory phase. However, a notable surge in publications occurred in 2020 and 2021, signaling a heightened academic interest in Click Fraud Detection as a major challenge within the digital advertising industry.

From a geographical perspective, the study demonstrates that India leads in terms of the number of publications in this area, followed by China, Saudi Arabia, and the United States. India's dominance not only reflects the rapid expansion of its information technology and digital marketing sectors but also corresponds with the substantial growth of its digital advertising economy. The high prevalence of click fraud underscores the urgent need for more effective machine learning-based detection methods and helps explain the considerable academic focus on this issue in the region.

The network map analysis conducted in this study identified four major clusters within the field of Click Fraud Detection, reflecting the conceptual evolution and interconnections in the research domain. These clusters not only represent distinct conceptual foci within the literature but also offer substantial practical implications for the digital advertising ecosystem.

- The first cluster, centered on cybersecurity, indicates that click fraud is frequently integrated with broader digital threats such as botnets and phishing, necessitating the development of ML-based mitigation strategies that can be embedded into the IT security infrastructures of advertising firms.
- The second cluster, which explores the relationship between click fraud and digital advertising business models such as real-time bidding (RTB), has direct implications for financial risk management, campaign budget optimization, and the selection of more credible publishing partners.
- The third cluster underscores the critical role of data processing techniques and model evaluation in addressing challenges such as class imbalance (class skewness), which is common in digital advertising datasets and can

degrade the performance of detection models. This highlights the need for advertising service providers to invest in machine learning systems capable of handling real-world data in a more representative manner.

- The fourth cluster, which focuses on system development and user behavior, emphasizes the importance of building adaptive detection architectures based on behavioral profiling. Such systems can be integrated into advertising platforms to monitor user activity in real-time and identify suspicious clicking patterns.

The identified thematic clusters reveal not only the current structure of research in click fraud detection but also highlight unresolved challenges that require sustained scholarly attention. For instance, the emergence of class skewness and SMOTE in Cluster 3 points to a persistent data imbalance problem that undermines model generalization—particularly when fraudulent instances represent a small fraction of total user behavior. This is a real-world constraint in ad ecosystems, where genuine traffic far exceeds malicious activity. Addressing this issue will require future research to move beyond oversampling techniques toward advanced solutions such as cost-sensitive learning, meta-learning, and anomaly-aware classifiers optimized for rare-event detection.

Similarly, the prominence of botnet, malware, and traffic pattern in Cluster 1 signals the growing sophistication of automated fraud actors that mimic legitimate click behavior. These trends call for detection systems that integrate behavioral profiling and network anomaly detection, capable of adapting to adversarial tactics in real time. Meanwhile, Cluster 4's focus on system, behavior, and architecture highlights a translation gap between algorithmic models and their deployment in production environments. This emphasizes the need for scalable, explainable models that can operate within latency-sensitive systems such as real-time bidding (RTB) platforms.

As a whole, these trends suggest that future research must be multidimensional: advancing algorithmic resilience, integrating domain-specific behavioral cues, and aligning model performance with operational constraints. A promising direction includes the development of end-to-end fraud detection pipelines that fuse unsupervised anomaly detection, explainable AI (XAI), and economic impact modeling, thereby enabling fraud mitigation strategies that are not only accurate but also actionable and transparent in commercial advertising environments. For instance, explainable AI frameworks such as LIME introduced for model-agnostic interpretability in general classification tasks [20] have since been widely adopted across domains requiring transparency and trust, including fraud detection and high-stakes automated decision-making. These developments underscore the growing feasibility of integrating transparency, adaptability, and interpretability into real-time fraud mitigation pipelines.

The overlay visualization in Fig 6 reveals a shift in focus within Click Fraud Detection research using Machine Learning. Keywords such as "performance evaluation" and "dataset", appearing in the yellow spectrum, indicate a growing emphasis on model evaluation and dataset quality in recent studies. This trend suggests that the scientific community is becoming increasingly aware of the importance of proper data management to enhance the performance of Click Fraud Detection models.

Additionally, the term "class skewness" has emerged as one of the high co-occurrence keywords in recent studies. This indicates that challenges related to class imbalance in datasets are becoming a major concern, as Click Fraud datasets typically exhibit an imbalanced class distribution [4], [5], [21], [22], [23], [24], [25], [26], [27]. Consequently, Machine Learning methods need to be adapted to effectively handle this issue.

IV. CONCLUSIONS

The bibliometric analysis in this study reveals that research on Click Fraud Detection has grown significantly in recent years, as evidenced by the increasing number of publications and the broadening scope of international collaboration. The keyword network mapping indicates that research in this domain can be categorized into four major clusters: cybersecurity, the digital advertising industry, dataset evaluation and processing, and the development of more adaptive detection systems. Recent research trends have shifted toward improving dataset quality and model evaluation, suggesting that data validity and the effectiveness of detection methods are becoming central concerns in current scholarly investigations.

Based on these findings, this study offers several practical recommendations. For researchers, it is essential to develop click fraud detection models that can operate in real-time and to implement approaches based on explainable AI (XAI) in order to enhance the transparency and accountability of detection systems. Furthermore, future research agendas should include the exploration of blockchain technology as a foundation for building more secure and decentralized digital advertising systems.

For regulators and policymakers, there is a need for stricter regulations regarding ad traffic verification, as well as the development of policy frameworks that facilitate ethical data sharing between digital platforms and research institutions. For the digital advertising industry, the adoption of real-time detection systems based on machine learning and behavioral profiling can strengthen resilience against click fraud manipulation while simultaneously improving the efficiency of advertising budget management.

ACKNOWLEDGMENT

The authors would like to thank the Buddhi Dharma University and the supervisors who have supported this research.

REFERENCES

[1] D. S. Soemarwoto, "PEMANTAPAN EKONOMI DIGITAL GUNA MENINGKATKAN KETAHANAN NASIONAL," *Jurnal Lembaga Ketahanan Nasional Republik Indonesia*, vol. 8, no. 1, pp. 1–6, Oct. 2022, doi: https://doi.org/10.55960/jlri.v8i1.299.

[2] W. Wang *et al.*, "Digital economy sectors are key CO2 transmission centers in the economic system," *J Clean Prod*, vol. 407, 2023, doi: 10.1016/j.jclepro.2023.136873.

[3] C. M. Simamora, R. Ningsih, P. Pendidikan, P. Perdagangan, P. Pengkajian, and P. L. Negeri, "INKLUSIVITAS EKONOMI DIGITAL DI INDONESIA: PERSPEKTIF GENDER DAN PENCIPTAAN LAPANGAN KERJA (STUDI KASUS KAMPUNG MARKETER)," 2020

[4] G. S. Thejas, S. Dheeshjith, S. S. Iyengar, N. R. Sunitha, and P. Badrinath, "A hybrid and effective learning approach for Click Fraud detection," *Machine Learning with Applications*, vol. 3, p. 100016, Mar. 2021, doi: 10.1016/j.mlwa.2020.100016

[5] M. Aljabri and R. M. A. Mohammad, "Click fraud detection for online advertising using machine learning," *Egyptian Informatics Journal*, vol. 24, no. 2, pp. 341–350, Jul. 2023, doi: 10.1016/j.eij.2023.05.006.

[6] D. Sisodia and D. S. Sisodia, "Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset," *Engineering Science and Technology, an International Journal*, vol. 28, Apr. 2022, doi: 10.1016/j.jestch.2021.05.015.

[7] M. Aljabri and R. M. A. Mohammad, "Click fraud detection for online advertising using machine learning," *Egyptian Informatics Journal*, vol. 24, no. 2, pp. 341–350, Jul. 2023, doi: 10.1016/j.eij.2023.05.006.

[8] R. Oentaryo *et al.*, "Detecting Click Fraud in Online Advertising: A Data Mining Approach Ghim-Eng Yap," *Journal of Machine Learning Research*, vol. 15, pp. 99–140, 2014, [Online]. Available: http://palanteer.sis.smu.edu.sg/fdma2012/.

[9] L. F. Cardona, J. A. Guzmán-Luna, and J. A. Restrepo-Carmona, "Bibliometric Analysis of the Machine Learning Applications in Fraud Detection on Crowdfunding Platforms," Aug. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jrfm17080352

[10] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *J Bus Res*, vol. 133, pp. 285–296, Sep. 2021, doi: 10.1016/j.jbusres.2021.04.070.

[11] W. Glänzel and H. F. Moed, "Opinion paper: Thoughts and facts on bibliometric indicators," *Scientometrics*, vol. 96, no. 1, pp. 381–394, 2013, doi: 10.1007/s11192-012-0898-z.

[12] J. C. Donohue, *Understanding Scientific Literatures: A Bibliometric Approach*. Cambridge, MA: MIT Press, 1974. [Online]. Available: https://mitpress.mit.edu/9780262040396/understanding-scientific-literatures/

[13] D. Guo, H. Chen, R. Long, H. Lu, and Q. Long, "A co-word analysis of organizational constraints for maintaining sustainability," *Sustainability (Switzerland)*, vol. 9, no. 10, pp. 1–19, Oct. 2017, doi: 10.3390/su9101928.

[14] A. Lis, "Keywords Co-occurrence Analysis of Research on Sustainable Enterprise and Sustainable Organisation," *Journal of CorporAte Responsibility and Leadership*, vol. 5, pp. 48–66, 2018, doi: 10.12775/JCRL.2018.011.

[15] A. Klarin, "How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews," *Int J Consum Stud*, vol. 48, no. 2, Mar. 2024, doi: 10.1111/ijcs.13031

[16] S. Sadeghpour and N. Vlajic, "Click fraud in digital advertising: A comprehensive survey," *Computers*, vol. 10, no. 12, Dec. 2021, doi: 10.3390/computers10120164.

[17] S. Nagaraja and R. Shah, "Clicktok: Click fraud detection using traffic analysis," in *WiSec 2019 - Proceedings of the 2019 Conference on Security and Privacy in Wireless and Mobile Networks*, Association for Computing Machinery, Inc, May 2019, pp. 105–116. doi: 10.1145/3317549.3323407.

[18] T. G.S., S. Dheeshjith, S. S. Iyengar, N. R. Sunitha, and P. Badrinath, "A hybrid and effective learning approach for Click Fraud detection," *Machine Learning with Applications*, vol. 3, p. 100016, Mar. 2021, doi: 10.1016/j.mlwa.2020.100016.

[19] Neeraja, Anupam, Sriram, S. Shaik, and V. Kakulapati, "Fraud Detection of AD Clicks Using Machine Learning Techniques," *J Sci Res Rep*, vol. 29, no. 7, pp. 84–89, Jun. 2023, doi: 10.9734/jsrr/2023/v29i71762.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[21] A. Batool and Y. C. Byun, "An Ensemble Architecture Based on Deep Learning Model for Click Fraud Detection in Pay-Per-Click Advertisement Campaign," *IEEE Access*, vol. 10, pp. 113410–113426, 2022, doi: 10.1109/ACCESS.2022.3211528.

[22] Y. Mhaske, A. Gupta, V. Bhosale, and K. Nair, "Click Fraud Detection Of Advertisements using Machine Learning," *International Research Journal of Engineering and Technology*, vol. 09, pp. 908–912, Apr. 2022, [Online]. Available: www.irjet.net

[23] Neeraja, Anupam, Sriram, S. Shaik, and V. Kakulapati, "Fraud Detection of AD Clicks Using Machine Learning Techniques," *J Sci Res Rep*, vol. 29, no. 7, pp. 84–89, Jun. 2023, doi: 10.9734/jsrr/2023/v29i71762.

[24] D. Sisodia and D. S. Sisodia, "A transfer learning framework towards identifying behavioral changes of fraudulent publishers in pay-per-click model of online advertising for click fraud detection," *Expert Syst Appl*, vol. 232, p. 120922, Dec. 2023, doi: 10.1016/j.eswa.2023.120922.

[25] D. Sisodia and D. S. Sisodia, "Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset," *Engineering Science and Technology, an International Journal*, vol. 28, p. 101011, Apr. 2022, doi: 10.1016/j.jestch.2021.05.015

[26] D. Sisodia and D. S. Sisodia, "Feature space transformation of user-clicks and deep transfer learning framework for fraudulent publisher detection in online advertising," *Appl Soft Comput*, vol. 125, p. 109142, Aug. 2022, doi: 10.1016/j.asoc.2022.109142

[27] F. Zhu, C. Zhang, Z. Zheng, and S. Al Otaibi, "Click Fraud Detection of Online Advertising-LSH Based Tensor Recovery Mechanism," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9747–9754, Jul. 2022, doi: 10.1109/TITS.2021.3107373.