

INTEGRATED FORENSICS DATA COLLECTION AND ANALYSIS MODEL FOR DATABASES USING DESIGN SCIENCE METHOD

*Ahmad Alshammari **

Department of Computer Sciences, Faculty of Computing and Information
Technology, Northern Border University, Rafha 91911
Kingdom of Saudi Arabia

* Corresponding Author, e-mail: Ahmad.Almkhaidsh@nbu.edu.sa

Abstract: Given the availability of database evidence, it is essential to implement effective forensic processes to maintain data integrity. Consequently, collecting and preserving digital evidence from the victim's database is crucial for analysis and presentation. This evidence must be gathered and examined in a forensically sound manner, using appropriate methodologies and tools. Although various collection and analysis models have been proposed based on existing literature, many are tailored to specific scenarios and database types. As a result, there is no universal, standardized model that is free from tailored approaches, but rather one that is suited for future database investigations. Thus, this paper introduces the Integrated Forensics Data Collection and Analysis Model for Databases (IFDCAMD), a unified framework designed to address the lack of standardization and redundancy in existing database forensic investigation processes. The model integrates key phases, including acquisition, preservation, reconstruction, analysis, and documentation & reporting, into a structured, forensically sound methodology. It aims to streamline investigations and enhance the admissibility of digital evidence, with applicability to post-incident response scenarios. The model's effectiveness is demonstrated through real-world case studies involving both relational (MySQL) and NoSQL (MongoDB) databases.

Key words: forensics of databases, investigation of databases, digital forensics, and investigative processes.

1. INTRODUCTION

In the domain of database forensics, its importance and uniqueness are attributable to the fact that it offers digital forensic investigators the opportunity to employ scientifically established methods for identifying, collecting, preserving, reconstructing, analyzing, and documenting database incidents, thereby facilitating the development of their digital forensic evidence [1]. The domain of database forensics, while somewhat complex, ambiguous, and diverse, is still regarded as a highly intricate and variable field because database systems are multidimensional [2]. Owing to their complexity and

heterogeneity, the database forensic domain depends on specific, repetitive collection and analysis models, which some perceive as unnecessary redundancies.

A database forensic investigation endeavors to gather digital evidence from the target database for purposes of analysis, preservation, reconstruction, and presentation in court as supporting evidence to substantiate claims and facts. Frequently, these investigations result in the formulation of hypotheses that can be employed in litigation and, in numerous instances, assist in proving a point in court.

This paper aims to highlight the importance of a coordinated approach that prioritizes collection and analysis as initial steps to reduce redundancy, addressing a research gap identified in this study. The authors introduce a harmonized model for defining the processes involved in collecting and analyzing evidence in database forensics, called IFDCAMD. This model stands out from existing frameworks through its integration of reconstruction, analysis, and collection, as well as its focus on post-incident investigation.

The processes outlined in this paper, however, are designed to address the need for innovative process models in post-incident responses, enabling database forensic investigations within the context of post-incident response strategies. Although this study is based on the initial proposed models that helped identify the research gap, it is essential to acknowledge that it relies on these models. In the following section, we will elaborate on this dependence in greater detail.

The IFDCAMD is a unified framework designed to resolve issues of standardization and redundancy in database forensic investigations. Unlike existing models, which are typically isolated, customized for specific scenarios, or restricted to a single type of database, the IFDCAMD provides a unified and standardized approach. Its principal feature is the seamless integration of five essential phases: Acquisition, Preservation, Reconstruction, Analysis, and Documentation & Reporting into a single, forensically sound process applicable to both relational and NoSQL databases.

The model's innovation lies in its applicability to both relational (MySQL) and NoSQL (MongoDB) databases, its comprehensive approach that merges database and system-level evidence, and its focus on producing legally admissible reports. This addresses a significant gap left by earlier models, which were often scenario-specific or limited to specific database management systems.

This research advances the field of digital forensics by specifically concentrating on the subdomain of database forensic investigation processes. It addresses a recognized research deficiency: the absence of a unified and standardized model for collecting and analyzing forensic data across various database technologies.

This paper is organized into six main sections: an Introduction that establishes the research gap in database forensics; Related Works that reviews and critiques existing forensic models; Methodology which details the Design Science Research approach used to develop the IFDCAMD model; Implementation where the model is validated through real-world case study involving MySQL and MongoDB; Findings and Discussion that analyses the model's effectiveness and compares its advantages to prior works; and a Conclusion that summarizes the achievement and suggests directions for future research.

2. RELATED WORKS

This section provides a critical analysis of the database forensics literature, showing that current models and processes are highly fragmented. They can be divided into two main, often separate, research streams.

a) **Models Focused on Evidence Collection and Preservation**

A significant portion of research concentrates on the early phases of an investigation. These studies recommend specific methods for gathering volatile and non-volatile artifacts; however, they are frequently tailored to a particular Database Management System (DBMS) or scenario. For instance, the author in [3] described artifact collection for MSSQL Server, while the authors in [4] suggested a file collection process for Oracle. Khanuja & Adane [5] and others [6-10], [11-13] provided similar frameworks, but each was created for a specific context (e.g., MySQL, fraud detection, and metadata extraction).

b) **Models Focused on Reconstruction and Analysis**

Another important focus area involves the later stages of analysing collected data. These models use various techniques to piece together events and detect malicious activity. For example, the authors in [4] used an algorithm to reconstruct intruder activity from backups. The authors in [18] highlighted ideal log settings for reconstruction, and authors in [20] concentrated on ‘internal structure carving.’ Other methods include media analysis [17], and financial data analysis for fraud [7].

c) **Comprehensive Models and Recent Contributions**

Some studies have pursued more comprehensive approaches, for example, the authors in [25] developed a unified incident response model aligned with ISO/IEC standards, and the authors in [26] designed a tamper detection model for NoSQL databases. Surveys, such as the one authors [21] thoroughly review the field; however, do not suggest a unified process.

Upon this critical review, three critical gaps are apparent:

1. *Lack of Integration*: The current models operate in silos. Processes for collection (Stream 1) and analysis/reconstruction (Stream 2) are often proposed separately, leading to potential disconnects, redundancy, and a lack of a unified end-to-end approach.
2. *Scenario and database management system specificity*: Many proposed models lack universality. They are tailored for databases, such as Oracle, MySQL, or MSSQL, or for specific incident types, such as financial fraud, which restricts their broader applicability.
3. *Absence of a Standardized, Phased Framework*: There is a notable lack of a single, structured, and forensically sound framework that combines all key phases from Acquisition to Documentation & Reporting into a unified, streamlined process suitable for various modern database systems (both relational and NoSQL).

As a result, the findings of this review directly motivate the current research. The IFDCAMD is proposed explicitly in this study to address these gaps by unifying previously disconnected processes into a single, universal, and standardized model.

3. METHODOLOGY

To underpin the theoretical assumptions and attain the primary objective of this study, the authors meticulously customized structured Design Science Research (DSR) methodologies to ensure they are both practical and efficient [29]. This careful adaptation is essential for accurately capturing knowledge in forensic science. The main reason for choosing DSR is its applicability; it involves defining a forensic study area, sampling it, and mapping the region to gather the necessary data to address the research gaps.

Therefore, the IFDCAMD consists of five main phases: Acquisition, Preservation, Reconstruction, Analysis, and Documentation & Reporting. The model is designed to unify and streamline forensic investigation processes for databases. Below is a summary of each component, along with a flowchart illustrating the method as shown in Figure 1.

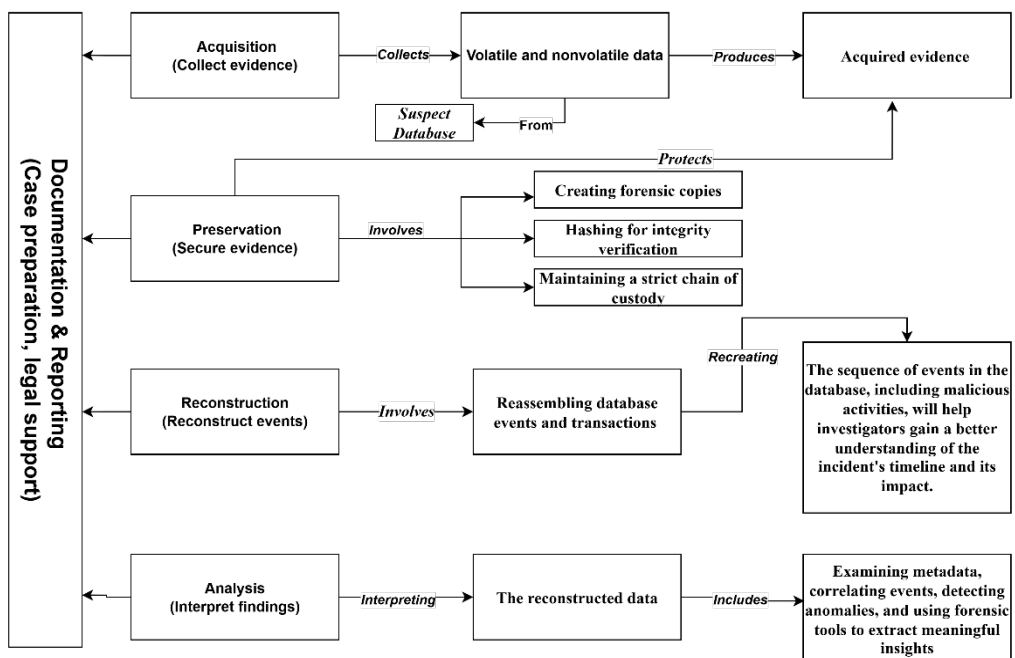


Figure 1. The developed IFDCAMD

- 1- **Acquisition:** The acquisition phase involves the systematic collection of both volatile and non-volatile artifacts from a suspect database. This includes data files, log files, transaction logs, data caches, and metadata. The goal is to capture a comprehensive snapshot of the database state at the time of the incident, ensuring that all potential evidence is gathered without alteration. This phase is critical for maintaining the integrity and admissibility of evidence in subsequent forensic analysis.
- 2- **Preservation:** Preservation ensures that the acquired evidence remains unaltered and tamper-proof throughout the investigation. This involves creating forensic copies, hashing for integrity verification, and maintaining a strict chain of custody. The phase is designed to protect the evidence from contamination or loss, thereby upholding its

validity for legal proceedings and ensuring that the original data remains pristine for further examination.

- 3- **Reconstruction:** Reconstruction involves reassembling database events and transactions from the collected artifacts, such as redo logs, undo logs, and transaction records. This phase aims to recreate the sequence of actions performed on the database, including malicious activities, to understand the incident's timeline and impact. Reconstruction helps investigators visualize data manipulations and identify unauthorized changes or accesses.
- 4- **Analysis:** The analysis phase involves interpreting the reconstructed data to identify malicious activities, validate hypotheses, and conclude. This includes examining metadata, correlating events, detecting anomalies, and using forensic tools to extract meaningful insights. The results are documented to support legal cases, provide actionable intelligence, and contribute to the overall understanding of the security incident.
- 5- **Documentation & Reporting:** This phase is the critical final stage where all investigative actions, findings, and evidence are formally recorded to create a comprehensive report suitable for legal proceedings and stakeholder communication. This phase ensures the integrity and admissibility of the evidence by meticulously documenting the entire chain of custody, the tools and methodologies used in each phase (Acquisition, Preservation, Reconstruction, Analysis), and the examiner's conclusions drawn from the reconstructed timeline of events. The resulting report translates complex technical processes and findings into a clear, structured, and authoritative document that presents the evidence to support or refute a hypothesis, ultimately enabling informed decision-making in court or within an organization.

4. IMPLEMENTATION

This section evaluates the effectiveness of the developed IFDCAMD in real-world scenarios and case study for

Case Study: Investigation of a Data Breach in a MongoDB NoSQL Database

Scenario: A healthcare application storing patient metadata in a MongoDB database reported a potential data breach. Suspicion arose from a sudden spike in outbound network traffic from the database server. The investigation aimed to confirm the breach, identify the extracted data, and determine the method of attack. The IFDCAMD model was adapted to the document-oriented nature of MongoDB. The investigation flow is illustrated in Figure 2.

- **Acquisition & Preservation:** Investigators acquired the MongoDB data files (WiredTiger storage engine files), the journal logs (journal/ directory), and system logs. The journal logs, which record all write operations, were particularly crucial. These were preserved with cryptographic hashes to maintain their evidential value.
- **Reconstruction:** Unlike SQL databases, MongoDB does not have a single transaction log in the same way. Reconstruction involved parsing the journal logs and the oplog (if replication was enabled) to reassemble the sequence of operations. Furthermore, network flow data was used to reconstruct the timeline of the data exfiltration.

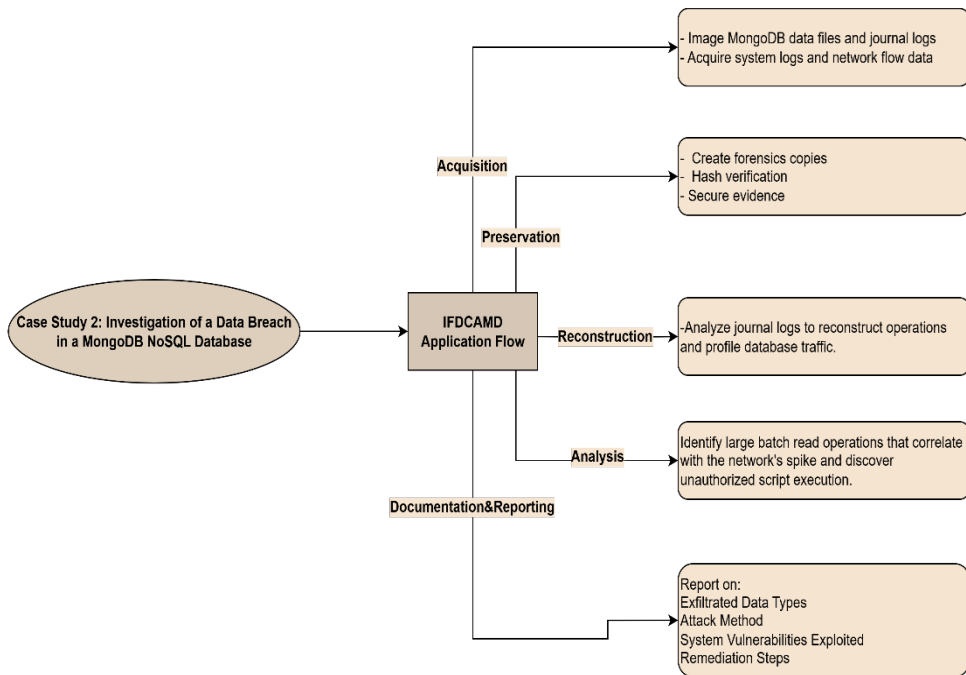


Figure 2. Case Study: Investigation of a Data Breach in a MongoDB NoSQL Database

- **Analysis:** The analysis phase focused on the journal logs, which revealed a series of large getMore operations (used for cursor iteration) on a collection containing patient data. These operations were timed precisely with the spike in outbound network traffic. Further investigation of system audit logs (if enabled) and process histories revealed an automated script that was executed via a compromised application programming interface (API) key, rather than a direct database login.
- **Documentation and Reporting:** The documentation and reporting phase produced a report critical for both technical remediation and regulatory compliance. The report meticulously documented the forensic process adapted for the NoSQL environment, highlighting how journal logs and network flow data were analysed to correlate massive batch read operations with the exfiltration of patient metadata to an external IP. It precisely identified the volume and type of compromised data and conclusively determined that the attack vector was a compromised API key, not a direct database breach. The report included all necessary elements for a data breach notification, such as the scope of impacted data and the root cause analysis, directing stakeholders toward addressing application-layer API security vulnerabilities. This thorough documentation provided a clear roadmap for incident response, ensuring compliance with potential legal and regulatory obligations.

The IFDCAMD model proved highly effective even in a NoSQL environment. Its phased structure forced a holistic view, combining database-specific artifacts (journal logs) with system-level evidence (network flows, process logs). The model successfully guided the investigation to confirm the breach, quantify the stolen data, and identify the

weak API authentication as the root cause, rather than a direct database vulnerability. This led to targeted remediation efforts.

5. DISCUSSION AND ANALYSIS

The development and deployment of the Integrated IFDCAMD requires a detailed discussion and evaluation of its effectiveness, advantages, and contributions to the field of database forensics. This evaluation utilizes case studies involving two database paradigms: relational databases (MySQL) and NoSQL databases (MongoDB). The focus is on how IFDCAMD effectively addresses the initial research gap: the absence of a standardized, universal framework that reduces redundancy and provides a reliable, forensically sound method for post-incident analysis.

A key focus of analysis is the model's integrated and phased design. Traditional database forensic models often treat steps like acquisition, preservation, reconstruction, analysis, and reporting as separate or sequential processes, which can lead to redundancy, oversight, and a lack of cohesion. IFDCAMD's innovation is its structure as a unified, coherent workflow where each phase logically transitions into the next, supporting a thorough and efficient investigation. The acquisition phase is notable for its systematic method of collecting both volatile and non-volatile artifacts, providing a complete snapshot of the database's state. This meticulous approach ensures that no evidence is missed from the start. The preservation phase's emphasis on cryptographic hashing and chain of custody is critically important because it maintains the integrity and legal admissibility of the evidence, which is essential for courtroom proceedings.

The reconstruction and analysis phases are highlighted as key areas where IFDCAMD demonstrates significant progress compared to earlier models. The approach treats reconstruction not as a separate technical task but as a crucial step to enable meaningful analysis. By assembling transaction sequences from logs like redo/undo logs in SQL, journal logs, and oplog in MongoDB, investigators can build an accurate timeline of events. This timeline forms the basis for the analysis phase, during which malicious patterns, anomalies, and specific intruder actions are identified and verified. The MongoDB case study illustrates this well; by linking parsed journal logs that show large getMore operations with network flow data, investigators not only confirmed the data breach but also precisely measured the extent of data exfiltration and pinpointed the exact incident time. This integration of database-level artifacts with system-level evidence, including network and process logs, is a key strength of the model, allowing for a comprehensive view of incidents that narrower models often miss.

Moreover, the analysis should consider the model's flexibility and broad applicability. Its successful use with both MySQL's structured schema and MongoDB's flexible, document-based system shows that IFDCAMD is not confined to a single database type. Its phases are generally applicable but can be tailored with technology-specific tools and methods. For instance, with MongoDB, the model directed investigators to examine WiredTiger storage files and journal logs instead of conventional SQL transaction logs. This adaptability directly addresses criticisms that existing models are often scenario-dependent or limited to a specific DBMS. IFDCAMD

is designed as a versatile meta-framework that can adjust to evolving database technologies, including future developments.

The discussion compares this work with the related studies discussed in Section 2. Unlike models that focus solely on artifact collection (e.g., [5, 10, 11]) or those dedicated to analysis and reconstruction (e.g., [4, 18, 20]), IFDCAMD integrates these steps. It also addresses a gap in other comprehensive models by explicitly including a robust Documentation & Reporting phase. This phase isn't just administrative; it's a crucial forensic step that turns technical findings into a clear, organized, and authoritative report suitable for legal and stakeholder review. This ensures that everything, from the chain of custody to the analytical results, is transparent and reproducible, thereby increasing the credibility and impact of the investigation's outcomes.

Therefore, the analysis would acknowledge the model's strategic value in post-incident response. By identifying the root cause of the MongoDB breach as a compromised API key rather than a direct database vulnerability, the investigation guided remediation efforts away from fruitless database hardening and toward crucial application-layer security fixes. This outcome underscores how the model's thoroughness leads to actionable intelligence and more effective cybersecurity resilience, moving beyond mere evidence collection to provide a clear roadmap for preventing future incidents. In conclusion, the discussion positions IFDCAMD as a validated, effective, and much-needed standardized framework that reduces redundancy, enhances evidence admissibility, and provides a structured methodology applicable across the diverse field of database systems.

6. CONCLUSION

The paper successfully developed and validated the Integrated Forensics Data Collection and Analysis Model for Databases (IFDCAMD), which provides a universal, structured, and forensically sound methodology that effectively unifies the previously fragmented and redundant processes of database forensic investigation, as demonstrated through its successful application in real-world case studies involving both relational (MySQL) and NoSQL (MongoDB) databases. For future work, the authors propose several considerations to further enhance the model, which likely include its formal validation across a broader range of database systems and incident types, the development of specialized tools to automate its phases, and its integration with broader incident response frameworks to improve overall cybersecurity resilience.

ACKNOWLEDGEMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work through the project number "NBU-FFR-2025-2990-08.

REFERENCES

- [1] Hungwe, Taurai, H. S Venter, V. R KEBANDE. Scenario-based digital forensic investigation of compromised MySQL database. *2019 IST-Africa Week Conference (IST-Africa)*, IEEE, 2019, pp. 1–11.

- [2] Al-Dhaqm, Arafat, S. Abd Razak, S. H. Othman, A. Ali, F. A. Ghaleb, A.S. Rosman, N. Marni. Database forensic investigation process models: A review. *IEEE Access*, vol. 8, 2020, pp. 48477–48490. doi: 10.1109/ACCESS.2020.2976885.
- [3] D. Litchfield. Oracle forensics part 4: Live response. *NGSSoftware Insight Secur. Res. (NISR), Next Gener. Secur. Softw. Ltd., Sutt.*, 2007.
- [4] Tripathi, S., B.B. Meshram. Digital evidence for database tamper detection. 2012.
- [5] Khanuja, H. K., D S Adane. A framework for database forensic analysis. *Computer Science Engineering*, vol. 2, no. 3, 2017, p. 27.
- [6] Lee, D., J. Choi, S. Lee. Database forensic investigation based on table relationship analysis techniques. *2009 2nd International Conference on Computer Science and Its Applications, CSA 2009*, 2009, p. 5404235.
- [7] Choi, J., K. Choi, S. Lee. Evidence investigation methodologies for detecting financial fraud based on forensic accounting. *2009 2nd International Conference on Computer Science and Its Applications, CSA 2009*, 2009, p. 5404202.
- [8] Olivier, M. S. On metadata context in database forensics. *Digital Investigation.*, vol. 5, no. 3–4, 2009, pp. 115–123.
- [9] Son, N., Keun-gi Lee, SangJun Jeon, H. Chung, S. Lee, Ch. Lee. The method of database server detection and investigation in the enterprise environment. *FTRA International Conference on Secure and Trust Computing, Data Management, and Application*, Springer, 2011, pp. 164–171.
- [10] Fowler, K. *SQL Server forensics analysis*. Pearson Education, 2008.
- [11] Ogutu, J. O. A methodology to test the richness of forensic evidence of database storage engine: Analysis Of MySQL update operation in InnoDB and MyISAM storageEngines. *University of Nairobi*, 2016/.
- [12] Khanuja, H. K., D. Adane. Forensic analysis of databases by combining multiple evidences. *International Journal of Computer Technology*, vol. 7, no. 3, 2013, pp. 654–663.
- [13] Frühwirt, P. K., K. Krombholz, E. Weippl. Towards a forensic-aware database solution: Using a secured database replication protocol and transaction management for digital investigations. *Digital Investigation*, vol. 11, no. 4, 2014, pp. 336–348.
- [14] Mohammed, Sh., R. Sridevi. A survey on digital forensics phases, tools and challenges. *Proceedings of the 3rd International Conference on Computational Intelligence and Informatics*, Springer, 2020, pp. 237–248.
- [15] Wong, D., K. Edwards. System and method for investigating a data operation performed on a database. Dec. 2005, *Google Patents*.
- [16] Susaimanickam, R. A workflow to support forensic database analysis. 2012, *Murdoch University*.
- [17] Fowler, K. A real-world scenario of a SQL Server 2005 database forensics investigation. *Information Security Reading room Pap. SANS Inst.*, 2007.
- [18] Adedayo, O.M., M. S. Olivier. Ideal log setting for database forensics reconstruction. *Digital Investigation*, vol. 12, 2015, pp. 27–40.

- [19] Khanuja, H., Sh.S. Suratkar. Role of metadata in forensic analysis of database attacks. *2014 IEEE International Advance Computing Conference (IACC)*, IEEE, pp. 457–462.
- [20] Wagner, J., Al. Rasin, J. Grier. Database forensic analysis through internal structure carving. *Digital Investigation*, vol. 14, 2015, pp. S106–S115.
- [21] Chopade, R., V.K. Pachghare. Ten years of critical review on database forensics research. *Digital Investigation*, vol. 29, 2019, pp. 180–197.
- [22] Orosco, Ch., C. Varol, N. Shashidhar. Graphically display database transactions to enhance database forensics. *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, 2020, pp. 1–6.
- [23] Adamu, B.Z., M. Karabatak, F. Ertam. A conceptual framework for database anti-forensics impact Mitigation,” in *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, 2020, pp. 1–6.
- [24] Marsh, R., S. Belguith, T. Dargahi. IoT database forensics: an investigation on HarperDB Security. *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems*, 2019, pp. 1–7.
- [25] Al-Dhaqm, A., Sh. Razak, S.H. Othman, A. Ngadi, M.N. Ahmed, A.A. Mohammed. Development and validation of a database forensic metamodel (DBFM). *PLoS One*, vol. 12, no. 2, 2017, doi: 10.1371/journal.pone.0170793.
- [26] Chopade, R., V. Pachghare. Data tamper detection from NoSQL Database in forensic environment. *Journal of Cyber Security. Mobility*, vol. 10, no. 2, 2021, pp. 421–450.
- [27] Al-Dhaqm, A., Sh. Razak, R.A. Ikuesan, V.R. Kebande, S.H. Othman. Face validation of database forensic investigation metamodel. *Infrastructures*, vol. 6, no. 2, 2021, doi: 10.3390/infrastructures6020013.
- [28] Choi, H., S. Lee, D. Jeong. Forensic recovery of SQL server database: Practical approach. *IEEE Access*, vol. 9, 2021, pp. 14564–14575.
- [29] Peffers, K. et al. A design science research methodology for information systems research,” *J. Manag. Inf. Syst.*, vol. 24, no. 3, 2007, pp. 45–77, 2007.

Information about the authors:

AHMAD ALSHAMMARI - received the M.S. degree in Computer Science from Saint Mary's University, USA, and Ph.D. degree in Computer Science from Oakland University, USA. He is currently working as an Associate Professor of Computer Science with the College of Computing and Information Technology, Northern Border University, Saudi Arabia. His research interests include information security, machine learning, data sciences, and IoT.

Manuscript received on 07 September 2025