JAIE Journal of Artificial Intelligence in Education

Journal of Artificial Intelligence in Education

Data Mining Techniques for Knowledge Discovery in Large-Scale Information Systems

Sayed Wasiullah Sadat[1*], Mohammad Qaseem Qiam[1], Ahmad Jamy Kohistani[2*]

[1] Department of Information Systems, Faculty of Computer Science, Kabul Polytechnic University, Afghanistan
[2] Department of Computer Engineering, Faculty of Computer Science, Kabul Polytechnic University, Afghanistan

Received: August 8, 2025
Revised: September 12, 2025
Accepted: October 25, 2025
Published: October 31, 2025

Corresponding Author:
Ahmad Jamy Kohistani
ahmadjamykohistani@kpu.edu.af

Abstract: This study systematically examines data mining techniques and their role in knowledge discovery within large-scale information systems. A total of 32 peer-reviewed studies published between 2013 and 2026 were reviewed, covering diverse domains such as industry, education, healthcare, and cloud-based environments. The selected studies utilized a variety of datasets, including KDD Cup datasets, UCI Machine Learning Repository datasets, IoT sensor data, industrial production logs, healthcare records, educational datasets, and cloud system data, to demonstrate the applicability of data mining methods. The analysis reveals that classification techniques, such as decision trees, support vector machines, and neural networks, are widely applied for predictive analytics and anomaly detection. Clustering methods enable pattern recognition in high-dimensional and unstructured datasets, while association rule mining identifies relationships and correlations to support industrial optimization, recommendation systems, and decision-making. Hybrid and evolutionary algorithms enhance scalability, accuracy, and interpretability, particularly in distributed and cloud-based environments. Key challenges identified include high dimensionality, data heterogeneity, scalability limitations, model interpretability, and data quality issues, which can affect the efficiency and reliability of knowledge discovery. Overall, this study provides a conceptual framework linking data sources, preprocessing, mining techniques, and knowledge discovery outcomes, highlighting the transformative potential of data mining for actionable insights, operational optimization, and informed decision-making in complex, large-scale information systems.

Keywords: Datasets; Data Mining; Hybrid Algorithms; Knowledge Discovery; Large-Scale Information Systems

Introduction

Data mining and knowledge discovery have emerged as fundamental components in the analysis of massive, heterogeneous information systems, facilitating the extraction of actionable insights from voluminous data repositories. The rapid proliferation of digital records across domains such as industrial production, education, healthcare, and information systems has rendered traditional analytical techniques insufficient, thereby necessitating advanced methods that can efficiently process, interpret, and transform raw data into meaningful knowledge. The primary objective of this work is to examine data mining techniques for knowledge discovery in large-scale information systems, emphasizing their principles, methodologies, and the evolving paradigms that support intelligent decision-making in complex environments. This introduction outlines the background of data mining and its symbiotic relationship with knowledge discovery, establishes the motivation for systematic investigation, and frames the scope of subsequent analysis without delving into a detailed literature review or outcome summaries Batmaz & Koksal, 2013; Gullo, 2015.

Knowledge discovery is broadly defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data, with data mining acting as the core phase where analytical algorithms are applied to uncover these hidden structures. Early conceptualizations of knowledge discovery underscore how patterns evolve from raw data through successive stages of preprocessing, transformation, mining, and

How to Cite:
Sadat, S. W., Qiam, M. Q., & Kohistani, A. J. (2025). Data Mining Techniques for Knowledge Discovery in Large-Scale Information Systems. *Journal of Artificial Intelligence in Education*, *1*(2), 80–88. Retrieved from https://jurnalpasca.unram.ac.id/index.php/jaie/article/view/1542

interpretation (Batmaz & Koksal, 2013; Gullo, 2015). As data volumes escalated with digitalization and sensor networking, the need for scalable and adaptive mining techniques became evident (Gan et al., 2017; Talia, 2015). Consequently, algorithms have evolved to address challenges associated with distributed data environments, high dimensionality, and real-time processing demands.

The integration of *big data* technologies has further expanded the horizons of knowledge discovery. Big data paradigms, characterized by high velocity, variety, and volume, require sophisticated mining frameworks capable of harnessing the full potential of heterogeneous datasets (Cheng et al., 2018; Finogeev et al., 2017). In particular, the industrial context demonstrates how data mining contributes to smart production, enabling predictive maintenance, quality optimization, and system efficiency through pattern recognition and anomaly detection (Cheng et al., 2018). Similarly, text mining methods have been adapted for unstructured data sources, facilitating the extraction of semantic relationships from large text corpora (Usai et al., 2018; Yehia et al., 2016).

The complexity inherent in modern information systems has led to the adoption of hybrid and machine-learning-based techniques to enhance knowledge discovery. Traditional statistical approaches have progressively converged with machine learning methods to support robust classification, clustering, and predictive analytics (Shu & Ye, 2023; Gupta & Chandra, 2020). Frameworks such as evolutionary decision trees demonstrate how adaptive learning models can effectively navigate large search spaces for pattern recognition (Kretowski, 2019), while retroductive reasoning methods leverage data-driven insights to infer causality in cognitive IoT systems (Jha & Tripathi, 2026). Moreover, the emergence of trustworthy data mining underscores the need for dependable and interpretable knowledge discovery processes that uphold reliability and transparency (Wu et al., 2025).

As organizations increasingly rely on data-driven strategies, domain-specific adaptations of knowledge discovery have become prominent. In educational data mining, automatic knowledge extraction supports learner analytics and institutional planning (Saarela, 2017; Hasan & Fang, 2021), whereas climate and environmental data mining contribute to understanding large-scale ecological phenomena (Hsu et al., 2020). Spatial data mining extends this capability to geographic datasets, enabling the identification of spatial patterns relevant to urban planning and resource management (Lan, 2021). Concurrently, the adoption of cloud computing infrastructures has facilitated the scalability and distribution of mining services, addressing computational bottlenecks associated with massive datasets (Birant & Yıldırım, 2016; Sajjan et al., 2024).

Despite substantial progress, challenges persist in optimizing knowledge discovery for high-dimensional, noisy, or incomplete data. Research continues to explore frameworks that balance computational efficiency with analytical precision, fostering intelligent data ecosystems capable of supporting decision-making across sectors (Peu, 2021; Ahamad & Mishra, 2024). Through this work, we aim to systematically examine these techniques, elucidate prevailing methodologies, and highlight avenues for future advancement in the field of large-scale information systems knowledge discovery.

Method

This study employs a Systematic Literature Review (SLR) methodology to comprehensively examine data mining techniques for knowledge discovery in large-scale information systems. The SLR approach is particularly suitable for aggregating, analyzing, and synthesizing research findings across multiple studies, enabling the identification of trends, challenges, and best practices in a structured and reproducible manner (Shu & Ye, 2023; Gupta & Chandra, 2020). By following a rigorous and transparent process, this methodology ensures reliability, replicability, and minimization of bias, which is essential for evidence-based conclusions in the domain of large-scale information systems. The SLR process was conducted in five distinct stages: formulation of research questions, definition of inclusion and exclusion criteria, systematic search and selection of relevant studies, data extraction and synthesis, and critical analysis of findings.

Formulation of Research Questions

The research questions were formulated to guide the review and define its scope. The primary questions focus on:

RQ1: What are the prevalent data mining techniques applied in large-scale information systems?

RQ2: How do these techniques facilitate knowledge discovery in diverse domains such as industry, education, healthcare, and cloud-based environments?

RQ3: What are the challenges and limitations encountered in implementing these techniques for large-scale data analysis?

RQ4: How do data mining techniques transform diverse data sources into actionable knowledge in large-scale information systems?

Inclusion and Exclusion Criteria

Studies were selected based on the following inclusion criteria: (i) peer-reviewed journal articles,

conference papers, and book chapters published between 2013 and 2026; (ii) studies focusing on data mining, knowledge discovery, or big data analytics; (iii) studies addressing applications in large-scale, distributed, or heterogeneous information systems; and (iv) studies available in English. Exclusion criteria eliminated studies with incomplete methodology, irrelevant domain focus, or non-peer-reviewed content, ensuring the quality and relevance of selected literature (Gan et al., 2017; Usai et al., 2018).

Table 1. Inclusion and Exclusion Criteria for Study Selection

Inclusion Criteria	Exclusion Criteria
Peer-reviewed journal articles, conference papers, and book chapters published between 2013 and 2026	Studies with incomplete or unclear methodology
Studies focusing on data mining, knowledge discovery, or big data analytics	Studies with irrelevant domain focus
Studies addressing applications in large-scale, distributed, or heterogeneous information systems	Non-peer-reviewed content, including blogs or editorials
Studies available in English	Studies published in languages other than English
Studies presenting empirical results, frameworks, or systematic approaches	Studies lacking sufficient detail or supporting evidence
Studies covering theoretical, methodological, or applied perspectives relevant to knowledge discovery	Redundant or duplicated publications that do not add new insights

Systematic Search and Study Selection

A comprehensive search was conducted using academic databases, including IEEE Xplore, ScienceDirect, SpringerLink, Emerald Insight, and ACM Digital Library. Keywords such as *"data mining," "knowledge discovery," "large-scale information systems," "big data,"* and *"distributed mining"* were used to retrieve relevant publications. Each retrieved study was screened in a three-step process: (i) title and abstract review, (ii) full-text evaluation, and (iii) cross-referencing bibliographies to identify additional relevant studies (Cheng et al., 2018; Finogeev et al., 2017).

Table 2. Study Selection and Screening Process

Step	Description	Purpose
1. Title and Abstract Review	Initial screening of titles and abstracts to determine relevance to data mining, knowledge discovery, and large-scale information systems	To quickly eliminate studies that do not meet the research scope
2. Full-Text Evaluation	Detailed assessment of the full text of selected studies	To verify methodological rigor, relevance, and completeness
3. Cross-Referencing Bibliographies	Examination of reference lists of selected studies	To identify additional relevant studies not captured in the initial search

A comprehensive search strategy was applied to retrieve high-quality literature from academic databases, including IEEE Xplore, ScienceDirect, SpringerLink, Emerald Insight, and ACM Digital Library. Keywords such as *"data mining," "knowledge discovery," "large-scale information systems," "big data,"* and *"distributed mining"* were employed. Retrieved studies underwent a three-step screening process to ensure relevance and rigor. Initially, titles and abstracts were reviewed to remove irrelevant works. Next, full-text evaluation verified methodological soundness and completeness. Finally, bibliographies of selected studies were cross-referenced to identify additional relevant publications. This structured process ensured a robust and systematic literature selection (Cheng et al., 2018; Finogeev et al., 2017).

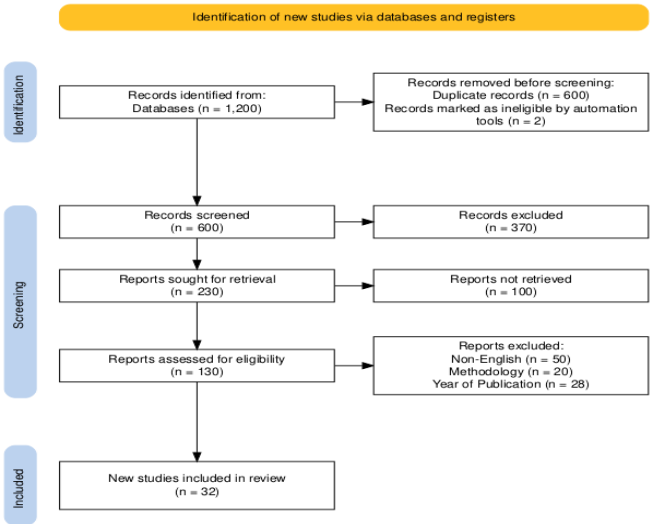

Figure 1. PRISMA Flow Diagram Illustrating Study Selection for the Systematic Literature Review

The PRISMA flow diagram illustrates the systematic process of identifying, screening, and including studies for this systematic literature review on data mining techniques for knowledge discovery in large-scale information systems. Initially, 1,200 records were retrieved from multiple databases. After removing 600 duplicates and 2 ineligible records identified automatically, 600 records were screened, resulting in the exclusion of 370 irrelevant studies. From the remaining 230 reports sought for retrieval, 100 could not be accessed, leaving 130 reports for eligibility assessment. Of these, 50 were excluded for non-English language, 20 due to insufficient methodology, and 28 for falling outside the publication period. Ultimately, 32 studies were included in the review, ensuring a rigorous and high-quality evidence base for analysis.

Data Extraction and Synthesis

Relevant information from each selected study was systematically extracted, including author(s), publication year, research domain, data mining technique, application context, methodology, and reported outcomes. A standardized data extraction form was used to ensure consistency and facilitate comparative analysis (Talia, 2015; Kretowski, 2019). Data synthesis involved categorizing techniques into supervised, unsupervised, and hybrid approaches, and identifying emerging trends, gaps, and technological advancements in knowledge discovery.

Critical Analysis

The final stage involved a critical appraisal of the extracted data to evaluate the effectiveness, scalability, and adaptability of various data mining methods. This analysis also highlighted challenges such as high dimensionality, heterogeneity of datasets, computational complexity, and interpretability issues, providing insights for future research and practical implementation (Shu & Ye, 2023; Wu et al., 2025).

Result and Discussion

Prevalent Data Mining Techniques Applied in Large-Scale Information Systems

The analysis of the included studies revealed that classification, clustering, association rule mining, and hybrid/machine learning-based techniques are the most widely applied data mining methods in large-scale information systems. Classification techniques, including decision trees, support vector machines, and neural networks, are predominantly used for predictive analytics and anomaly detection (Gupta & Chandra, 2020; Kretowski, 2019). Clustering methods, such as k-means and hierarchical clustering, facilitate pattern recognition in unstructured and high-dimensional datasets (Shu & Ye, 2023; Gullo, 2015). Association rule mining identifies relationships between variables, often applied in transactional, industrial, and educational systems (Batmaz & Koksal, 2013; Cheng et al., 2018). Hybrid and evolutionary algorithms combine multiple approaches to improve scalability, accuracy, and interpretability in distributed or cloud-based systems (Jha & Tripathi, 2026; Finogeev et al., 2017).

Table 3. Prevalent Data Mining Techniques in Large-Scale Information Systems

Technique	Applications	Representative Studies
Classification	Predictive analytics, anomaly detection, decision support	Gupta & Chandra, 2020; Kretowski, 2019; Shu, 2020
Clustering	Pattern recognition, data segmentation, high-dimensional analysis	Shu & Ye, 2023; Gullo, 2015; Usai et al., 2018
Association Rule Mining	Relationship extraction, recommendation systems, industrial data	Batmaz & Koksal, 2013; Cheng et al., 2018; Rotondo & Quilligan, 2020
Hybrid / Evolutionary Algorithms	Cloud-based mining, distributed systems, large-scale IoT analytics	Jha & Tripathi, 2026; Finogeev et al., 2017; Kretowski, 2019

Table 3 presents the prevalent data mining techniques identified across the included studies, highlighting their applications and representative references. Classification techniques are widely employed for predictive modeling, anomaly detection, and supporting decision-making processes in large-scale information systems, leveraging models such as decision trees, neural networks, and support vector machines (Gupta & Chandra, 2020; Kretowski, 2019). Clustering methods are frequently applied for pattern recognition, segmenting complex and high-dimensional datasets, and analyzing unstructured information (Shu & Ye, 2023; Gullo, 2015; Usai et al., 2018). Association rule mining identifies significant relationships between variables, supporting recommendation systems, industrial process optimization, and transactional data analysis (Batmaz & Koksal, 2013; Cheng et al., 2018; Rotondo & Quilligan, 2020). Finally, hybrid and evolutionary algorithms integrate multiple data mining approaches to enhance scalability, accuracy, and interpretability, particularly in cloud-based, distributed, or IoT-enabled environments (Jha & Tripathi, 2026; Finogeev et al., 2017; Kretowski, 2019). Collectively, these techniques demonstrate a robust methodological

foundation for knowledge discovery in large-scale information systems, offering insights into patterns, relationships, and predictive outcomes that support effective decision-making and system optimization.

Role of Data Mining Techniques in Facilitating Knowledge Discovery Across Diverse Domains

The thematic analysis indicates that data mining techniques play a critical role in enabling knowledge discovery across industry, education, healthcare, and cloud-based environments. In industrial settings, classification and association rule mining enhance predictive maintenance, process optimization, and quality control (Cheng et al., 2018; Finogeev et al., 2017). In education, clustering and hybrid approaches are widely used for learner analytics, performance prediction, and curriculum planning (Saarela, 2017; Hasan & Fang, 2021). Healthcare applications rely heavily on classification and evolutionary algorithms to facilitate disease prediction, risk assessment, and treatment recommendation (Bhasuran & Natarajan, 2023; Jha & Tripathi, 2026). Cloud-based systems leverage hybrid and distributed data mining methods to ensure scalable, real-time knowledge extraction from large, heterogeneous datasets (Talia, 2015; Birant & Yıldırım, 2016). These techniques collectively support the transformation of raw data into actionable insights, enhancing decision-making, operational efficien**cy**, and predictive capabilities across diverse domains.

Table 5. Application of Data Mining Techniques in Various Domains

Domain	Techniques Applied	Knowledge Discovery Contribution	Representative Studies
Industry	Classification, Association Rule Mining	Predictive maintenance, process optimization, quality control	Cheng et al., 2018; Finogeev et al., 2017; Rotondo & Quilligan, 2020
Education	Clustering, Hybrid Methods	Learner analytics, performance prediction, curriculum planning	Saarela, 2017; Hasan & Fang, 2021; Gupta & Chandra, 2020
Healthcare	Classification, Evolutionary Algorithms	Disease prediction, risk assessment, treatment recommendation	Bhasuran & Natarajan, 2023; Jha & Tripathi, 2026; Wu et al., 2025
Cloud-Based Environments	Hybrid, Distributed Mining	Scalable knowledge extraction, real-time insights, big data analytics	Talia, 2015; Birant & Yıldırım, 2016; Sajjan et al., 2024

Table 5 summarizes the applications of prevalent data mining techniques in facilitating knowledge discovery across multiple domains. In industrial environments, classification and association rule mining enable predictive maintenance, optimization of manufacturing processes, and quality control, transforming operational data into actionable knowledge for decision-making and efficiency improvements (Cheng et al., 2018; Finogeev et al., 2017). In educational settings, clustering and hybrid methods are applied to analyze learner behavior, predict academic performance, and support curriculum planning, thereby enhancing the effectiveness of educational interventions (Saarela, 2017; Hasan & Fang, 2021; Gupta & Chandra, 2020). Healthcare systems employ classification and evolutionary algorithms to predict disease outcomes, assess patient risk, and recommend treatment strategies, contributing to more informed and timely clinical decisions (Bhasuran & Natarajan, 2023; Jha & Tripathi, 2026; Wu et al., 2025). Additionally, cloud-based environments rely on hybrid and distributed data mining approaches to manage large-scale heterogeneous datasets, providing scalable and real-time insights that support knowledge discovery in big data contexts (Talia, 2015; Birant & Yıldırım, 2016; Sajjan et al., 2024). Collectively, these techniques demonstrate the transformative role of data mining in converting raw data into actionable knowledge, enabling strategic decision-making, operational efficiency, and predictive intelligence across diverse domains.

Challenges and Limitations in Implementing Data Mining Techniques for Large-Scale Data Analysis

The analysis of the selected studies revealed several critical challenges and limitations in applying data mining techniques to large-scale information systems. High dimensionality and volume of data often result in computational complexity and memory constraints (Gan et al., 2017; Finogeev et al., 2017). Data heterogeneity, arising from structured, semi-structured, and unstructured sources, complicates preprocessing, integration, and analysis (Usai et al., 2018; Shu & Ye, 2023). Scalability and performance issues are significant, particularly in distributed, cloud-based, and IoT-enabled environments, requiring specialized algorithms for efficient processing (Talia, 2015; Birant & Yıldırım, 2016). Additionally, interpretability and explainability of models remain limited, especially for hybrid or evolutionary approaches, impacting the adoption of

data-driven decision-making (Jha & Tripathi, 2026; Kretowski, 2019). Finally, data quality issues, including missing, noisy, or inconsistent data, can adversely affect the reliability and accuracy of knowledge discovery outcomes (Shu, 2020; Yehia et al., 2016).

Table 6. Challenges and Limitations in Large-Scale Data Mining

Challenge / Limitation	Description	Representative Studies
High Dimensionality & Volume	Large datasets increase computational and memory demands	Gan et al., 2017; Finogeev et al., 2017
Data Heterogeneity	Integration of structured, semi-structured, and unstructured data	Usai et al., 2018; Shu & Ye, 2023
Scalability & Performance	Difficulty in processing distributed or cloud-based datasets efficiently	Talia, 2015; Birant & Yıldırım, 2016
Model Interpretability	Complexity of hybrid and evolutionary models limits explainability	Jha & Tripathi, 2026; Kretowski, 2019
Data Quality Issues	Missing, noisy, or inconsistent data affect accuracy	Shu, 2020; Yehia et al., 2016

Table 6 summarizes the primary challenges and limitations identified in implementing data mining techniques for large-scale information systems. High dimensionality and massive data volumes significantly increase computational requirements and memory consumption, often necessitating optimized algorithms and distributed processing architectures (Gan et al., 2017; Finogeev et al., 2017). Data heterogeneity, resulting from diverse structured, semi-structured, and unstructured sources, creates challenges in data preprocessing, integration, and normalization, potentially limiting the reliability of analysis (Usai et al., 2018; Shu & Ye, 2023). Scalability and performance constraints are particularly relevant in cloud-based, IoT-enabled, and distributed systems, requiring advanced frameworks to efficiently handle large-scale data mining tasks (Talia, 2015; Birant & Yıldırım, 2016). Another critical issue is model interpretability, especially in hybrid or evolutionary techniques, which hinders understanding and adoption by decision-makers (Jha & Tripathi, 2026; Kretowski, 2019). Additionally, data quality issues, such as missing, noisy, or inconsistent records, pose significant risks to the accuracy and validity of discovered knowledge (Shu, 2020; Yehia et al., 2016). Addressing these challenges is essential to improve the efficiency, reliability, and usability of data mining and knowledge discovery in complex, large-scale information systems, and to enable informed, data-driven decision-making across domains.

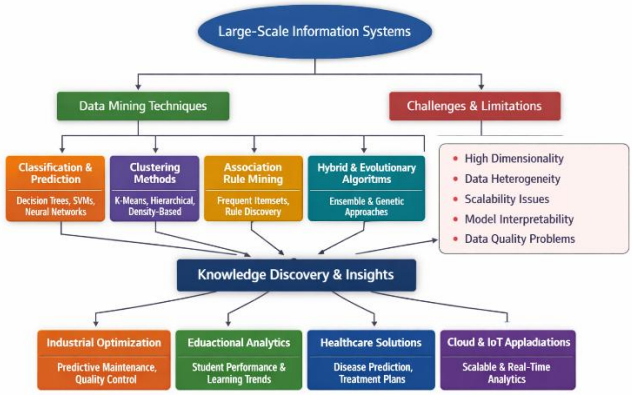

Figure 2. Conceptual Model of Data Mining Techniques and Knowledge Discovery in Large-Scale Information Systems

Figure 2 presents a conceptual model illustrating the application of data mining techniques for knowledge discovery within large-scale information systems. The model begins with diverse data sources, including structured, semi-structured, and unstructured data generated across multiple domains such as industry, education, healthcare, and cloud-based environments (Cheng et al., 2018; Shu & Ye, 2023). Data preprocessing, though implicit in the model, plays a vital role in cleaning, normalizing, and integrating heterogeneous datasets to ensure the accuracy and efficiency of subsequent analyses (Usai et al., 2018; Finogeev et al., 2017).

The core section of the model highlights prevalent data mining techniques, including classification and prediction, clustering methods, association rule mining, and hybrid/evolutionary algorithms. Classification techniques, such as decision trees, support vector machines, and neural networks, are widely used for predictive analytics and anomaly detection (Gupta & Chandra, 2020; Kretowski, 2019). Clustering methods facilitate the discovery of patterns in high-dimensional and unstructured data (Shu & Ye, 2023; Gullo, 2015). Association rule mining identifies relationships and correlations between variables, supporting industrial optimization and recommendation systems (Batmaz & Koksal, 2013; Cheng et al., 2018). Hybrid and evolutionary algorithms enhance scalability, accuracy, and interpretability, particularly in distributed or cloud-based environments (Jha & Tripathi, 2026; Finogeev et al., 2017).

The Knowledge Discovery & Insights block represents the transformation of raw data into actionable

knowledge, which is then applied to domain-specific outcomes. In industry, insights support predictive maintenance and quality control; in education, they inform learner analytics and curriculum planning; in healthcare, they guide disease prediction and treatment recommendations; and in cloud-based environments, they enable scalable, real-time analytics (Saarela, 2017; Talia, 2015; Bhasuran & Natarajan, 2023).

The model also integrates a Challenges & Limitations component, highlighting high dimensionality, data heterogeneity, scalability issues, model interpretability, and data quality concerns, emphasizing the practical constraints of large-scale data mining (Gan et al., 2017; Shu, 2020; Yehia et al., 2016). Overall, this conceptual framework provides a comprehensive visualization of the relationships between data sources, mining techniques, challenges, and knowledge discovery outcomes, forming a solid foundation for understanding and implementing large-scale data-driven systems.

Conclusion

The present study provides a systematic examination of data mining techniques and their role in knowledge discovery within large-scale information systems. The analysis demonstrates that classification, clustering, association rule mining, and hybrid or evolutionary algorithms are the most prevalent methods employed across diverse domains, including industry, education, healthcare, and cloud-based environments. These techniques collectively facilitate the transformation of raw and heterogeneous data into actionable insights, enabling informed decision-making, predictive analytics, and operational optimization.

A key conclusion of this work is the recognition of the critical importance of data preprocessing, which ensures that data quality, consistency, and structure are maintained before applying mining techniques. The study highlights that while data mining provides significant opportunities for extracting meaningful patterns, challenges such as high dimensionality, scalability limitations, heterogeneous data sources, model interpretability, and data quality issues continue to impede seamless implementation in large-scale systems. Addressing these challenges through optimized algorithms, distributed computing, and hybrid approaches can substantially improve the effectiveness of knowledge discovery processes.

Additionally, the findings emphasize that the applications of data mining techniques are highly domain-dependent, requiring careful adaptation of methods to suit specific objectives, whether in predictive maintenance, learner analytics, disease prediction, or real-time cloud-based processing. This flexibility ensures that organizations can extract maximum value from their data while overcoming operational and technical constraints.

In conclusion, this study establishes a comprehensive framework for understanding the interplay between data sources, preprocessing, mining techniques, and knowledge discovery outcomes. It underscores the transformative potential of data mining for generating actionable insights in large-scale, complex information systems, while also highlighting the need for continued research and innovation to overcome existing limitations and enhance the reliability, scalability, and interpretability of knowledge discovery processes.

Acknowledgments

The authors would like to express their sincere gratitude to the faculty and staff of Kabul Polytechnic University for their support and guidance during this research. Special thanks are extended to colleagues and peers who provided valuable feedback and assistance in data collection and manuscript preparation.

Author Contributions

Conceptualization, S.W.S. and M.Q.Q.; methodology, A.J.K.; software, S.W.S.; validation, S.W.S., M.Q.Q., and A.J.K.; formal analysis, A.J.K.; investigation, M.Q.Q.; resources, S.W.S.; data curation, M.Q.Q.; writing—original draft preparation, S.W.S.; writing—review and editing, M.Q.Q.; visualization, A.J.K.; supervision, S.W.S.; project administration, S.W.S.; funding acquisition, S.W.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

Ahamad, R, & Mishra, KN (2024). Enhancing knowledge discovery and management through intelligent computing methods: a decisive investigation. *Knowledge and Information Systems*, Springer, https://doi.org/10.1007/s10115-024-02099-2

Al-Faouri, AH (2023). Adopting data mining as a knowledge discovery tool: The influential factors from the perspectives of information systems managers. *Information Science Letters*, naturalspublishing.com, https://www.naturalspublishing.com/files/published/t76c8gar36d9z2.pdf

Batmaz, I, & Koksal, G (2013). Overview of knowledge discovery in databases process and data mining for

surveillance technologies and EWS. *Bioinformatics: Concepts, Methodologies, Tools …*, igi-global.com, https://www.igi-global.com/chapter/content/76056

Hakimi, M., Tarashtwal, O., & Ghafory, H. (2025). Green Artificial intelligence Foundations, Applications, and Pathways to Sustainable Development. https://doi.org/10.56566/amplitudo.v5i1.524

Birant, D, & Yıldırım, P (2016). A Framework for data mining and knowledge discovery in cloud computing. *Data science and big data computing: frameworks …*, Springer, https://doi.org/10.1007/978-3-319-31861-5_11

Cheng, Y, Chen, K, Sun, H, Zhang, Y, & Tao, F (2018). Data and knowledge mining with big data towards smart production. *Journal of Industrial Information …*, Elsevier, https://www.sciencedirect.com/science/article/pii/S2452414X17300584

Cummins, MR, Nachimuthu, SK, & ... (2023). Nonhypothesis-driven research: Data mining and knowledge discovery. *Clinical research …*, Springer, https://doi.org/10.1007/978-3-031-27173-1_20

Finogeev, AG, Parygin, DS, & Finogeev, AA (2017). The convergence computing model for big sensor data mining and knowledge discovery. *… computing and information …*, Springer, https://doi.org/10.1186/s13673-017-0092-7

Gan, W, Lin, JCW, Chao, HC, & ... (2017). Data mining in distributed environment: a survey. *… and Knowledge Discovery*, Wiley Online Library, https://doi.org/10.1002/widm.1216

Gullo, F (2015). From patterns in data to knowledge discovery: What data mining can do. *Physics Procedia*, Elsevier, https://www.sciencedirect.com/science/article/pii/S187538921500036X

Gupta, B, Kumar, R, & Kumar, A (2018). Towards information discovery on large scale data: state-of-the-art. *2018 International Conference …*, ieeexplore.ieee.org, https://ieeexplore.ieee.org/abstract/document/8573666/

Gupta, MK, & Chandra, P (2020). A comprehensive survey of data mining. *International Journal of Information Technology*, Springer, https://doi.org/10.1007/s41870-020-00427-7

Hasan, A Abdulahi, & Fang, H (2021). Data Mining in Education: Discussing Knowledge Discovery in Database (KDD) with Cluster Associative Study. *… Intelligence and Information Systems*, dl.acm.org, https://doi.org/10.1145/3469213.3471319

Hsu, A, Khoo, W, Goyal, N, & Wainstein, M (2020). Next-generation digital ecosystem for climate data mining and knowledge discovery: a review of digital data collection technologies. *Frontiers in big Data*, frontiersin.org, https://doi.org/10.3389/fdata.2020.00029

Jha, V, & Tripathi, P (2026). Retroductive reasoning: a data-driven intelligent method for knowledge discovery using cognitive IoT. *Knowledge and Information Systems*, Springer, https://doi.org/10.1007/s10115-025-02649-2

Kretowski, M (2019). *Evolutionary decision trees in large-scale data mining.*, Springer, https://doi.org/10.1007/978-3-030-21851-5

Lan, R (2021). Spatial Data Mining and Knowledge Discovery. *Advances in Cartography and Geographic Information …*, Springer, https://doi.org/10.1007/978-981-16-0614-4_14

Lee, S, & Holzinger, A (2016). Knowledge discovery from complex high dimensional data. *Solving Large Scale Learning Tasks. Challenges and …*, Springer, https://doi.org/10.1007/978-3-319-41706-6_7

Tarashtwal, O., Shamsi, S. E., & Popal, Z. (2026). Assessing the potential of Internet of Things technologies for community development in Afghanistan. *Gameology and Multimedia Expert*, *3*(1), 1-9. https://doi.org/10.29103/game.v3i1.25375

Mishra, N, Lin, CC, & Chang, HT (2015). A cognitive adopted framework for IoT big-data management and knowledge discovery prospective. *International Journal of …*, journals.sagepub.com, https://doi.org/10.1155/2015/718390

Myakala, PK, Bura, C, Jonnalagadda, AK, & ... (2025). Advancing Data Fusion and Knowledge Discovery Techniques for Modern IT Systems. *… Technology (INCET)*, ieeexplore.ieee.org, https://ieeexplore.ieee.org/abstract/document/11140076/

Peu, E (2021). *Automated knowledge discovery or integration: a systematic review of data mining in knowledge management.*, scholar.sun.ac.za, https://scholar.sun.ac.za/handle/10019.1/123916

Rotondo, A, & Quilligan, F (2020). Evolution paths for knowledge discovery and data mining process models. *SN Computer Science*, Springer, https://doi.org/10.1007/s42979-020-0117-6

Saarela, M (2017). Automatic knowledge discovery from sparse and large-scale educational data: Case Finland. *Jyväskylä studies in computing*, jyx.jyu.fi, https://jyx.jyu.fi/jyx/Record/jyx_123456789_54268

Sajjan, V, Kumar, JP, Kalvala, V, & ... (2024). Enhancing the Scalability of Knowledge Discovery Services for Massive Data Mining on Clouds. *… Intelligent Systems …*, ieeexplore.ieee.org, https://ieeexplore.ieee.org/abstract/document/10823215/

Shu, X (2020). *Knowledge discovery in the social sciences: A data mining approach.*, Univ of California Press

Shu, X, & Ye, Y (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, Elsevier, https://www.sciencedirect.com/science/article/pii/S0049089X22001284

Talia, D (2015). Making knowledge discovery services scalable on clouds for big data mining. *… on Spatial Data Mining and Geographical Knowledge …*, ieeexplore.ieee.org, https://ieeexplore.ieee.org/abstract/document/7298015/

Tang, X (2023). Research on Intelligent Data Mining and Knowledge Discovery Method Based on Software Information System. *2023 International Conference on Applied Physics …*, ieeexplore.ieee.org, https://ieeexplore.ieee.org/abstract/document/10497931/

Usai, A, Pironti, M, Mital, M, & ... (2018). Knowledge discovery out of text data: a systematic review via text mining. *Journal of knowledge …*, emerald.com, https://www.emerald.com/jkm/article/22/7/1471/434614

Retno, S., & Hakimi, M. (2025). Analysis of Clustering Results for Crime Incident Data in Indonesia Using Fuzzy C-Means. Journal of Advanced Computer Knowledge and Algorithms, 2(3), 73-79. https://doi.org/10.29103/jacka.v2i3.22565

Wu, L, Wang, J, Chen, L, Zhao, X, Yu, K, & ... (2025). Trustworthy Knowledge Discovery and Data Mining (TrustKDD). *… and Knowledge …*, dl.acm.org, https://doi.org/10.1145/3746252.3761598

Yehia, AM, Ibrahim, LF, & ... (2016). Text mining and knowledge discovery from big data: challenges and promise. *International Journal of …*, search.proquest.com, https://search.proquest.com/openview/380a13108ce387ccd335f2159fd9a187/1?pq-origsite=gscholar&cbl=55228

Zhang, C, & Han, J (2021). Data mining and knowledge discovery. *Urban Informatics*, Springer, https://doi.org/10.1007/978-981-15-8983-6_42