A Survey on the Convergence of Big Data, Artificial Intelligence and Data Mining

Enlai Wang[1], Jingbo Liu[1], Yufei Zhang[1], Xue Liu[2]
[1]Henan University of Technology, Zhengzhou, Henan, China
[2]iFLYTEK Co., Ltd., Hefei, Anhui, China

Abstract: This paper provides a comprehensive study of the convergence of big data, Artificial Intelligence (AI), and data mining, systematically analyzing their inherent logic and arguing that these technologies form a closed-loop, value-creating ecosystem. Big data serves as the foundation, data mining as the bridge, and AI as the ultimate goal. The study proposes a layered technological architecture to support this synergy and explores its transformative applications in various fields such as finance, healthcare, and intelligent recommender systems. The research highlights the enormous potential for improving efficiency and fostering innovation while also identifying key challenges hindering widespread adoption, including data privacy issues, the "black box" nature of complex AI models, algorithmic bias, and ever-increasing computational demands. Finally, the paper discusses future trends, including Automated Machine Learning (AutoML), privacy-preserving computation, and explainable artificial intelligence (XAI), emphasizing the need to balance development processes to reconcile technological progress with ethical governance.

Keywords: Big Data; Artificial Intelligence; Data Mining; Technological Convergence; Intelligent Applications

1. Introduction

With the steady development of modern economy and technology, people's daily lives have undergone tremendous changes, and we are at a historical intersection driven by data. In the digital economy era, data, as a new type of key production factor, is growing in scale and complexity at an unprecedented speed, constituting the macro phenomenon we call "big data". The wave of digital economy and the dawn of the intelligent era together outline the future picture. Data is no longer just a carrier of information, but has become a key production factor on par with land, labor, capital and technology. Against this grand background, the ability to process and utilize massive amounts of data has become a key pillar for countries, industries and enterprises to build core competitiveness. The three engines driving this change—big data, data mining and artificial intelligence—are penetrating and deeply integrating with each other at an unprecedented speed, reshaping the way we perceive and transform the world. How to accurately and efficiently extract insights and knowledge from this ocean of information and transform them into intelligent decisions and actions has become the core challenge and opportunity facing academia and industry. This transformation process "from data to intelligence" is a vivid practice of the deep interweaving and integration of the three major technology fields of big data, data mining and artificial intelligence.

To understand this profound transformation, we must first clarify the inherent characteristics and division of labor of the three core concepts. Big data refers to massive datasets with large scale and complex and diverse structures. They are enormous in size, reaching PB or even EB levels; they are diverse in type, covering structured, semi-structured and unstructured text, images, videos, etc., and are the foundation and fuel. Their characteristics are usually summarized by the "4Vs": fast processing speed, requiring real-time or near-real-time analysis and response of data, extremely difficult to store, analyze and visualize, and difficult to use for subsequent processing or to obtain results [1,2], low value density, and the high-value information contained in massive data may only be the tip of the iceberg. Big data constitutes a new challenge and a new resource for us. Data mining refers to the use of a complex set of tools and algorithms by analysts and end users to solve data problems. Data mining can also be regarded as a set of tools that can help users discover patterns and trends based on specific subsets of enterprise data [3,4]. Data mining is a core technology and alchemy. It is a series of processes that systematically explore, filter, and extract previously unknown, potentially useful, and ultimately understandable hidden information and patterns from the "mine" of big data through machine learning and statistical algorithms (such as classification, clustering, association analysis, and anomaly detection). It is a key step in transforming raw data into "intelligent" raw materials.

There are many definitions of artificial intelligence. In the Turing Test, artificial intelligence is defined as the ability of a machine (using electronic output devices) to communicate with humans without revealing its non-human identity. Its basic judgment criteria are binary [5]. In the current situation of the integration of big data, artificial intelligence and data mining, artificial intelligence is positioned as the ultimate goal and intelligent engine. It aims to study and develop theories, methods and application systems for simulating, extending and expanding human intelligence. Its goal is to enable machines to perceive, learn, reason and make decisions like humans, and even surpass the boundaries of human capabilities. Big data provides learning materials for AI, while data mining provides methods for extracting knowledge from materials. The two together support the training and evolution of AI models.

In summary, the core argument of this paper is that the deep integration of big data, data mining, and artificial intelligence is no longer a simple technological aggregation, but rather forms a complete closed-loop paradigm of "data-driven intelligence." This paradigm is driving scientific research, socio-economic development, and various industries towards intelligent transformation with unprecedented breadth and depth. This paper will systematically outline the integration mechanism of these three technologies, comprehensively investigate their current applications in various fields, deeply analyze the key challenges currently faced, and provide a forward-looking perspective on future development trends.

2. Literature Review

With the rapid development of information technology, the integration of big data, artificial intelligence (AI) and data mining has become the core driving force for promoting the intelligent transformation of various fields of society. The three do not exist in isolation, but constitute a closely coupled technological ecosystem: big data provides the raw materials and foundation for analysis, data mining is the key means to extract knowledge and laws from it, and artificial intelligence uses this knowledge to build intelligent models, ultimately realizing the automation and intelligence of decision-making and application [6]. This section will sort out the current application status of this technology system in different fields based on existing literature.

In the agricultural field, which is of vital importance to the national economy and people's livelihood, the integration of the three is triggering a revolutionary change. Zhang et al. [7] pointed out that the next generation of artificial intelligence and big data-driven intelligent design is completely changing the traditional model of crop breeding. By integrating massive amounts of genomic, phenotypic and environmental data (big data), using advanced data mining technology to identify key genes and trait associations, and using AI models (such as deep learning) to predict phenotypes and optimize breeding programs, the breeding process of superior varieties has been greatly accelerated, demonstrating the huge potential of "intelligent design" in addressing global food security challenges. In traditional engineering fields such as geotechnical engineering, the integration technology is driving its transformation towards refinement and intelligence. Li [8] explored data mining and intelligent analysis of geotechnical engineering exploration based on artificial intelligence and big data. This study collected a large amount of geological exploration data (big data), applied data mining algorithms (such as clustering and classification) to summarize and analyze geotechnical parameters, and constructed an AI intelligent analysis model to achieve accurate evaluation and prediction of engineering geological conditions, effectively improving the reliability of engineering design and construction safety. In the field of business and information services, intelligent recommendation systems are one of the most typical applications of the integration of the three. Shi [9] studied the application of artificial intelligence-driven data mining algorithms in intelligent recommendation systems. Based on user behavior big data, the system uses data mining technology enhanced by AI algorithms (such as neural networks) to deeply mine users' potential interests and preference patterns, thereby achieving accurate information push and personalized services, significantly improving user experience and business conversion efficiency. In the field of public services, integrated technologies are becoming an important tool for improving the modernization of governance capabilities. Zhao [10] analyzed the logic, challenges and practices of artificial intelligence empowering university data governance. The study pointed out that universities can integrate multi-source heterogeneous data (big data) such as management, teaching and scientific research by building a data platform, use data mining technology to discover governance blind spots and optimization paths, and use AI capabilities to achieve intelligent decision support, ultimately promoting the transformation of education management from "experience-driven" to "data-driven". Similarly, Meng [11] also explored the application of AI and data mining in the construction of university library network reading service platforms, aiming to provide users with smarter knowledge services. Although the application prospects are broad, the deep integration of the three also faces common challenges. Arora et al. [6] emphasized that when conducting data mining and processing in the era of big data and artificial intelligence, it is essential to pay close attention to data privacy, security, and ethical issues. How to fully utilize the value of data while ensuring that personal privacy is not violated and that algorithmic decision-making is fair and transparent is a key issue that urgently needs to be addressed in future technological development and applications.

In summary, the integrated application of big data, artificial intelligence, and data mining has permeated numerous fields, including agricultural breeding, engineering construction, commercial services, and public governance, demonstrating a powerful enabling effect. Its core paradigm lies in using data mining as a bridge to transform the "quantity" of big data into the "quality" of artificial intelligence, ultimately achieving a leap in industry production and management efficiency. However, the healthy development of this technology still requires continued attention and resolution of the ensuing data security and ethical challenges.

3. The Inherent Logic and Technological System of Integration

The integration of big data, artificial intelligence, and data mining is not a simple accumulation of technologies, but rather an organic whole formed by their interdependence and mutual promotion. Understanding its inherent integration mechanism and outlining its technological architecture is key to grasping the development trajectory of this field.

3.1 Analysis of the Integration Mechanism

The integration of the three constitutes a closed-loop feedback system from data to intelligence, and Figure 1 shows its internal logic.

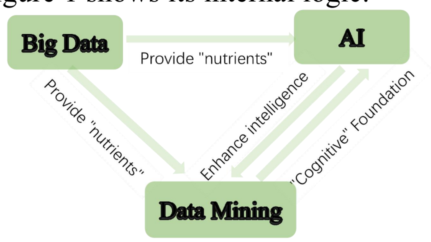

Figure 1. A Closed-Loop Feedback System from Data to Intelligence

(1) Big Data Provides "Nourishment" and a Testing Ground for AI and Data Mining

Massive, multi-source big data is the cornerstone of an integrated ecosystem. First, it provides a sufficient information foundation for solving complex problems, enabling data-driven deep learning models, in particular, to escape the predicament of "having no rice to cook." Second, the diversity of big data (such as text, images, and videos) has spurred the development of multimodal AI models, propelling artificial intelligence from perceptual intelligence to cognitive intelligence. Finally, the complex relationships and long-tail effects inherent in big data provide rich scenarios for model training, greatly enhancing the generalization ability and robustness of AI systems. Without big data, advanced AI models are like soldiers lacking practical training, unable to cope with the complexities of the real world.

(2) Data Mining is the "Perceptor" and "Cognitive" Foundation of AI

Data mining plays the role of extracting information and knowledge from raw data, and is a prerequisite for AI to achieve "intelligence." In AI systems, data mining techniques are widely used in feature engineering, pattern discovery, and association analysis. For example, clustering analysis can be used to group users, and association rule mining can be used to discover potential connections between products. These provide direct evidence for AI's accurate decision-making and recommendation. In this sense, data mining constitutes the "perception system" of AI, which enables AI to "understand" the preliminary patterns behind the data and lay the foundation for higher-level reasoning and decision-making.

(3) AI enhances the "intelligence" and automation of data mining

Artificial intelligence, especially its subfields such as natural language processing (NLP) and knowledge graphs, has profoundly transformed the traditional paradigm of data mining [12]. AI technology has enabled the data mining process to leap from "manual" to "automatic". For example, deep learning models can automatically extract high-dimensional features, replacing the tedious and experience-dependent manual feature engineering; reinforcement learning can automatically optimize the data mining process and parameter configuration; and NLP technology can directly and automatically mine key information from unstructured text reports. AI has injected "intelligent soul" into data mining, making it no longer a single, fixed algorithm tool, but an analysis system that can optimize itself and learn autonomously.

Big data is the foundation, data mining is the means, and artificial intelligence is the goal. Together, they form a spiraling value creation chain of "data → information → knowledge → intelligence".

3.2 Technical System

At the technical implementation level, the integration of the three forms a well-structured and collaborative technical system, the core integration diagram of which is shown in Figure 2.

The data processing and computing layer forms the underlying support of the integrated system, primarily responsible for the storage, preprocessing, and parallel computing of massive amounts of data. The core technologies of this layer include distributed file systems (such as HDFS), distributed computing frameworks (such as Spark and Flink), and resource management and scheduling systems (such as YARN and Kubernetes). Together, they construct a scalable and highly available data lake or data platform, providing a stable and efficient data pipeline for upper-layer data analysis and intelligent applications.

The analysis and modeling layer is the core engine of the integrated system, the main battleground where data mining algorithms and AI models converge and interact. In this layer, traditional machine learning algorithms (such as decision trees, SVM, and clustering algorithms) and cutting-edge deep learning models (such as CNN, RNN, and Transformer) work side-by-side. Data mining techniques are often used for initial data exploration and feature preprocessing, while AI models are responsible for complex pattern recognition and accurate prediction. Furthermore, the emergence of Automated Machine Learning (AutoML) and deep learning frameworks (such as TensorFlow and PyTorch) has significantly lowered the technical barrier to this layer, accelerating the value transformation process from data to models.

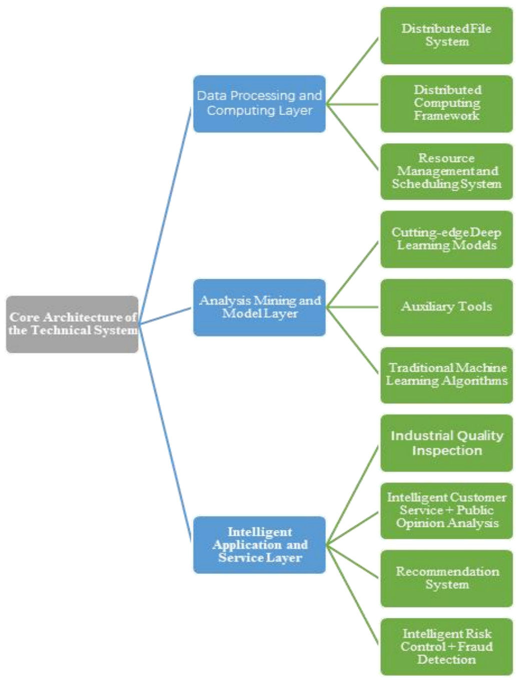

Figure 2. Core Technology Integration Diagram

The intelligent application and service layer represents the ultimate embodiment of integrated value, transforming the knowledge and models obtained from the lower layers into concrete business solutions. This layer, geared towards vertical industries, fosters a diverse range of intelligent applications, such as: industrial quality inspection based on computer vision, intelligent customer service and public opinion analysis based on natural language processing, recommendation systems based on collaborative filtering and deep learning, and intelligent risk control and fraud detection systems integrating knowledge graphs and data mining. These applications, through APIs, software services, or embedded systems, empower various industries with data intelligence, driving business innovation and efficiency improvements.

4. Scenarios and Common Challenges of Integrated Applications

The deep integration of big data, artificial intelligence, and data mining has spurred transformative applications in numerous fields. However, while appreciating its immense potential, we must be keenly aware that these practices face a series of common challenges in areas such as data, technology, and ethical governance. This chapter will analyze typical scenarios to reveal the dilemmas hidden beneath its apparent success.

4.1 Scenarios for Integrated Applications

In the financial sector, converged technologies have become a core pillar of risk prevention. The system first aggregates massive amounts of user transaction records, social network data, and behavioral logs at the big data level; then, it utilizes clustering and anomaly detection algorithms from data mining to identify fraudulent groups and abnormal patterns from these massive transactions; finally, a real-time risk control model built by artificial intelligence makes millisecond-level intelligent judgments on each transaction. In the healthcare field, converged technologies are driving healthcare towards precision and personalization. Applications rely on big data platforms to integrate multi-source medical imaging, electronic medical records, and genomics data; data mining techniques analyze the complex correlation rules between diseases and genes, and lifestyle habits; while artificial intelligence, especially deep learning models, has demonstrated expert-level capabilities in assisted diagnosis using CT/MRI images (such as tumor identification). Intelligent recommendation systems are one of the most widespread applications of this convergence. These systems use big data technology to record users' click, browsing, purchase, and rating history in real time; they use algorithms such as collaborative filtering and association rules from data mining to uncover the correlation between user preferences and items; and finally, artificial intelligence models (such as Wide & Deep) provide precise personalized recommendations.

4.2 Challenges and Dilemmas

Financial data is highly sensitive, posing significant security and privacy risks. Furthermore, data silos between different institutions make it difficult to create a complete risk profile, limiting model accuracy. Complex deep learning risk control models are often considered "black boxes" due to their poor interpretability, making it difficult for rejected loan applicants to understand the reasons, and posing challenges for financial institutions to regulatory oversight. If the training data itself contains implicit biases, the model can learn and amplify these algorithmic biases, potentially leading to systemic discrimination against specific genders, races, or geographic groups, raising serious questions about fairness.

The medical field may face challenges such as inconsistent quality of medical data. For example, unstructured medical records have low standardization, directly impacting model performance. Furthermore, data like genomics demands extremely high privacy protection, posing unprecedented security and privacy challenges. The "black box" problem of AI diagnostic models is particularly critical in this scenario, as doctors and patients find it difficult to trust a system that cannot provide diagnostic evidence. Simultaneously, processing massive amounts of bioinformatics data requires extremely high computing power, potentially exacerbating the "digital divide" in healthcare resource allocation in precision medicine. Moreover, data misuse (such as using genetic data for insurance pricing or employment discrimination) will bring unprecedented legal and ethical problems.

The continuous collection of user behavior data during intelligent recommendation processes has raised widespread concerns about privacy breaches. In an effort to capture subtle user interests, models have become increasingly complex, further exacerbating the interpretability dilemma and making it difficult for users to understand "why I saw this content." In pursuing maximum click-through rates, recommendation algorithms are prone to the "information cocoon" effect, solidifying and narrowing users' perspectives. This manipulation of user attention is itself a form of algorithmic bias and potential data abuse, which may negatively impact the construction of social consensus.

The above scenario analysis demonstrates that the integrated application of big data, artificial intelligence, and data mining has yielded remarkable results in improving industry efficiency and achieving precise services. However, technical challenges such as data quality and privacy, model interpretability, and computing power bottlenecks, as well as governance challenges such as algorithmic bias, data misuse, and lagging legal and ethical considerations, have become key bottlenecks restricting its healthy and sustainable development. In the future, promoting the development of technological integration must be carried out in parallel with building a responsible AI governance system to effectively mitigate its potential risks while enjoying the benefits of technology.

5. Future Development Trends

With the continuous iteration of information technology and the deepening of industrial digital transformation, the integrated system of big data, data mining, and artificial intelligence is evolving toward a more intelligent, secure, and standardized direction. Its future development trends can be summarized into three dimensions: technical evolution, application expansion, and ecological improvement.

5.1 Technical Trends

First, Automated Machine Learning (AutoML) will further lower the technical threshold for intelligent applications. By automating key processes such as feature engineering, model selection, hyperparameter tuning, and model deployment, AutoML enables non-professional technical personnel to quickly build high-quality AI models, promoting the popularization of intelligent technology in small and medium-sized enterprises and vertical segments. Second, privacy-preserving computing technologies represented by federated learning, homomorphic encryption, and differential privacy will become a core technical pillar for data value release under regulatory constraints. These technologies realize the goal of "data availability without visibility," allowing multiple parties to collaborate on model training without sharing original data, thus resolving the contradiction between data sharing and privacy protection. Third, Explainable AI (XAI) will gain increasing attention and application. As AI models are widely used in critical fields such as finance, medical care, and law, the demand for model transparency and interpretability is becoming more urgent. XAI technologies will break the "black box" limitation of traditional deep learning models, enabling users to understand the decision-making logic and reasoning process of models, thereby improving the trustworthiness and reliability of intelligent systems.

5.2 Application Trends

On one hand, the integration of intelligent technology with Internet of Things (IoT) and 5G will accelerate the realization of global intelligence. The large-scale deployment of IoT devices generates massive real-time data, while 5G technology provides high-speed, low-latency data transmission support. The combination of the two will promote the intelligent upgrade of scenarios such as smart cities, industrial Internet, and smart healthcare, realizing full-link intelligence from data collection, analysis, to decision-making. On the other hand, intelligent applications will shift from "point-based intelligence" to "systematic intelligence." In the early stage, intelligent technologies were often applied to single business links (such as single-user recommendation or partial quality inspection). In the future, they will form end-to-end intelligent systems covering the entire business chain, realizing collaborative optimization between different links and maximizing the value of intelligence. For example, in the manufacturing industry, intelligent systems will integrate production scheduling, quality control, logistics distribution, and after-sales service to achieve full-cycle intelligent management of products.

5.3 Ecological Trends

The data element market will gradually improve and mature, becoming a key driving force for the development of the intelligent industry. With the improvement of data classification, grading, and pricing mechanisms, data will be transformed into tradable production factors, promoting the compliant circulation and efficient allocation of data resources across regions and industries. At the same time, AI governance frameworks and relevant laws and regulations will continue to be improved. As AI technology brings potential risks such as algorithmic discrimination, data leakage, and job substitution, countries around the world will accelerate the formulation of relevant regulatory policies and ethical guidelines. This will not only restrict the disorderly development of technology but also provide a clear institutional guarantee for the healthy development of the intelligent industry, guiding technology to develop in the direction of benefiting humanity.

6. Conclusion

This research systematically expounds on the technical system, integration mechanism, and application value of the deep integration of big data, data mining, and artificial intelligence. The core conclusion of this research is that big data, data mining, and artificial intelligence have formed a positive cycle of "data-driven intelligence and intelligence creating value." Big data provides the foundation for intelligent analysis, data mining acts as a bridge connecting data and models, and artificial intelligence realizes the deep excavation of data value. This cycle not only promotes the continuous innovation of technology itself but also brings profound changes to various industries, improving production efficiency, optimizing service quality, and creating new business models.

In the future, while promoting the continuous development of intelligent technology, we must also attach great importance to the accompanying challenges. First, data security needs to be placed in a prominent position, and technical and institutional means should be used to build a solid barrier for data security. Second, ethical norms for AI should be strengthened to ensure that technology develops within the framework of morality and law and avoid adverse impacts on society and individuals. Third, talent training should be accelerated to cultivate interdisciplinary talents who master big data, data mining, and AI technologies, providing intellectual support for the sustainable development of the intelligent industry. Only by balancing technological progress and social responsibility can we promote the healthy and sustainable development of the intelligent society and better realize the goal of technology serving humanity.

References

[1] S Sagiroglu, D Sinanc. Big data: A review//2013 international conference on collaboration technologies and systems (CTS). IEEE, 2013: 42-47.

[2] Z Sun, Y Huo. The Spectrum of Big Data Analytics. Journal of Computer Information Systems, 2019, 61(2): 154-162. https://doi.org/10.1080/08874417.2019.1571456.

[3] Mining W I D. Introduction to data mining. Mining Multimedia Databases, Mining Time Series and, 2006.

[4] D Baylor, E Breck, H T Cheng, et al. Tfx: A tensorflow-based production-scale machine learning platform. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.

[5] Y Jiang, X Li, H Luo, et al. Quo vadis artificial intelligence? Discover Artificial Intelligence, 2022, 2(1): 4.

[6] S Arora, S R Thota, S Gupta. Data Mining and Processing in the Age of Big Data and Artificial Intelligence - Issues, Privacy, and Ethical Considerations//2024 4th Asian Conference on Innovation in Technology. 2024:1-6.

[7] Y Zhang, G M Huang, Y X Zhao, et al. Revolutionizing Crop Breeding: Next-Generation Artificial Intelligence and Big Data-Driven Intelligent Design. Engineering (English), 2025, 44(1): 245-255. DOI: 10.1016/j.eng.2024.11.034.

[8] S Y Li. Data Mining and Intelligent Analysis of Geotechnical Engineering Investigation Based on Artificial Intelligence and Big Data. Intelligent Building & Smart City, 2025(6): 47-49. DOI: 10.13655/j.cnki.ibci.2025.06.013.

[9]Y Y Shi. Research on the Application of Artificial Intelligence-Driven Data Mining Algorithms in Intelligent Recommendation

Systems. Software, 2025, 46(7): 70-72. DOI: 10.3969/j.issn.1003-6970.2025.07.021.
[10]L L Zhao. Artificial intelligence empowers data governance in universities: logic, challenges and practice. Chongqing Higher Education Research, 2022, 10(1): 71-79. DOI: 10.15998/j.cnki.issn1673-8012.2022.01.008.
[11]X D Meng. Application of artificial intelligence and data mining in the construction of online reading service platform for university libraries. Information Recording Materials, 2024, 25(9): 78-80.
[12]B Pan. Reconstruction of the rational use system of generative artificial intelligence text and data mining. Communication and Copyright, 2025(10): 107-110.
