# Big Data and Data Mining: Review of Concepts and Techniques

Aditya Kumar [1], Satya Pal [2], Abhishek Saxena [3], Maduresh kumar Yadav [4], Mohammad Jeelani [5]
[1,2] *Master of Computer Application, Faculty of Computer Application, Future University, Bareilly, India*
[3] *Dean, Faculty of Computer Application, Future University, Bareilly, India*
[4,5] *Associate Professor [4,5], Faculty of Computer Application, Future University, Bareilly, India*
doi.org/10.64643/IJIRTV13I3-207542-459

***Abstract*—The volume, velocity, and variety of data generated across digital platforms — social media, financial technology, scientific instrumentation, and networked sensors — have grown at an unprecedented pace, making Big Data and Data Mining two of the most actively studied areas in computer science today. This paper presents an updated review of the core concepts, data attribute types, and algorithmic techniques (decision trees, genetic algorithms, and artificial neural networks) that underpin modern data mining practice. Building on classical foundations, the review also situates these concepts within recent developments in machine learning and deep learning, and surveys current application domains including healthcare analytics, financial technology, cybersecurity/IoT intrusion detection, and cloud-based big data mining. In addition to summarizing techniques, this paper highlights the persistent challenges of scalability, data heterogeneity, and the gap between data-rich repositories and actionable knowledge ("information poverty"). Visual illustrations are used throughout to clarify the three V's of Big Data, the knowledge-discovery pipeline, and representative algorithmic architectures. The review concludes that while data mining methodology has matured considerably, opportunities for improvement remain wide open, particularly as data continues to grow in scale and complexity.**

***Index Terms*—data; big data; data mining; machine learning; deep learning; knowledge discovery The main purpose of data mining is classification or forecast. In classification, there are a lot of cases that can be simulated such as in a product ad can be sorted out where the respondents who are interested in the ad and which are not, from here can be developed about what is behind someone interested in the ad. this information is expected to be predicted through data mining.**

## I. INTRODUCTION

Since the digital era began and the internet began to be used extensively in the early 1990s, it has produced tremendous amounts of data transactions. Even long before the internet era of things a few decades earlier several studies had discussed the data of the washouse, big data and data mining.

Analogously, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. However, the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. [1]

Data mining is used to analyze and explore large amounts of data to find a valuable result from the extraction. there is a lot of information that can be extracted from a collection of databases that can be used for several needs, for example a company engaged in the commercial sector can utilize the transaction data to find optimal sales patterns. Thus, the income from a company can be increased through the use of data mining.

## II. DATA ATTRIBUTE

According to the type of data can be divided into:

### A. Nominal Attributes

Nominal means "relating to names." The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful

order. In computer science, the values are also known as enumerations.

B. **Binary Attributes**

A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as Boolean if the two states correspond to true and false.

C. **Ordinal Attributes**

An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.

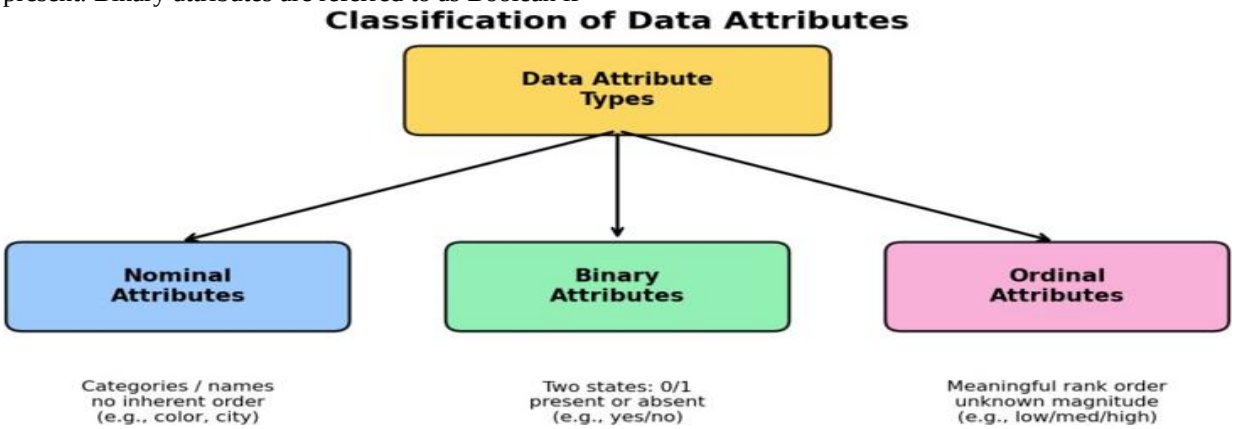


*Figure 1. Classification of data attributes into Nominal, Binary, and Ordinal types.*

## III. CONCEPTS

A. Big Data

Big data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity. This is known as the three Vs.

- Volume — The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a webpage or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.
- Velocity — Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.
- Variety — Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video require additional preprocessing to derive meaning and support metadata.

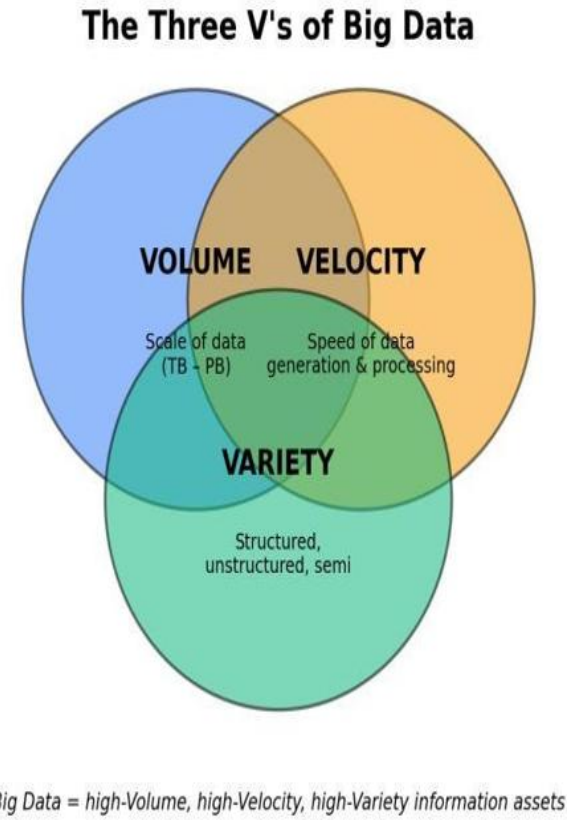


*Figure 2. The three V's of Big Data — Volume, Velocity, and Variety.*

B. Data Mining

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes.

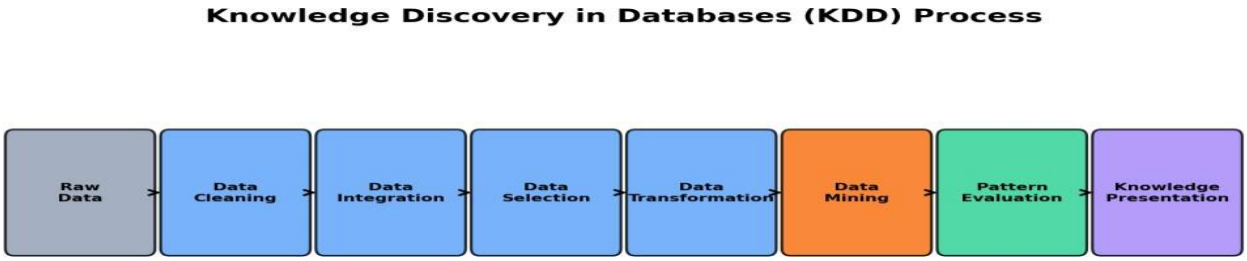


*Figure 3. The Knowledge Discovery in Databases (KDD) process, from raw data to actionable knowledge.*

## IV. ALGORITHM

A. Decision Tree Algorithm

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data). The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

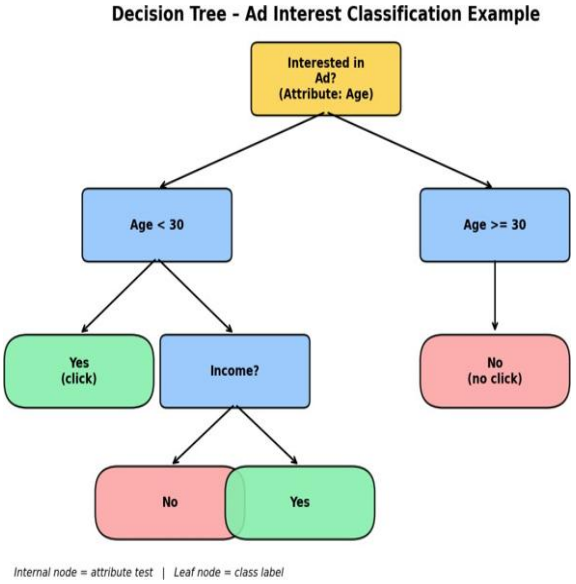


*Figure 4. A simplified decision tree for classifying interest in an advertisement.*

B. Genetic Algorithms (Complex Adaptive Systems)

A genetic or evolutionary algorithm applies the principles of evolution found in nature to the problem of finding an optimal solution to a Solver problem. In a "genetic algorithm," the problem is encoded in a series of bit strings that are manipulated by the algorithm; in an "evolutionary algorithm," the decision variables and problem functions are used directly. Most commercial Solver products are based on evolutionary algorithms. An evolutionary algorithm for optimization is different from "classical" optimization methods in several ways.

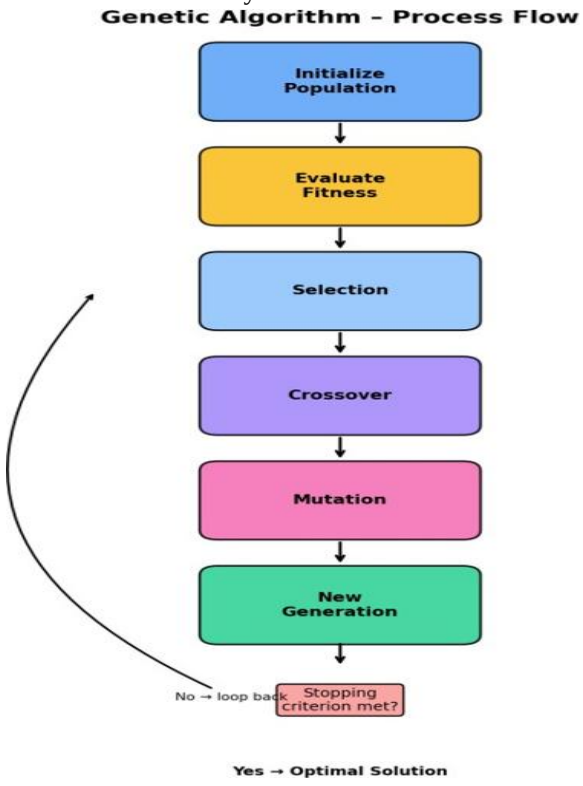


*Figure 5. Process flow of a genetic algorithm: initialization, fitness evaluation, selection, crossover, and mutation.*

C. Neural Networks

An artificial neural network is made up of a series of nodes. Nodes are connected in many ways like the neurons and axons in the human brain. These nodes are primed in a number of different ways. Some are limited to certain algorithms and tasks which they perform exclusively. In most cases, however, nodes are able to process a variety of algorithms. Nodes are able to absorb input and produce output. They are also connected to an artificial learning program. The program can change inputs as well as the weights for different nodes.

Weighting places more focus on different nodes or different functions of the same group of nodes. This network also has a function for displaying and evaluating the output. Evaluation is at the heart of the neural network and what makes it intelligent. Many types of computer systems run off of networks that have different nodes which perform different functions. However, weighting and artificial evaluation immensely increase the number of possible functions for the network.

Once it is established, the machine learning algorithm must be assigned and implemented. These networks need a basic type of function to be implemented by an operator. This function can be one of many different forms such as Apriori or k-means clustering. The artificial neural network then goes to work. It absorbs input and sends different tasks to different nodes. These nodes may work independently or they may frequently communicate with one another. The nodes then transfer an output to an end series of nodes.

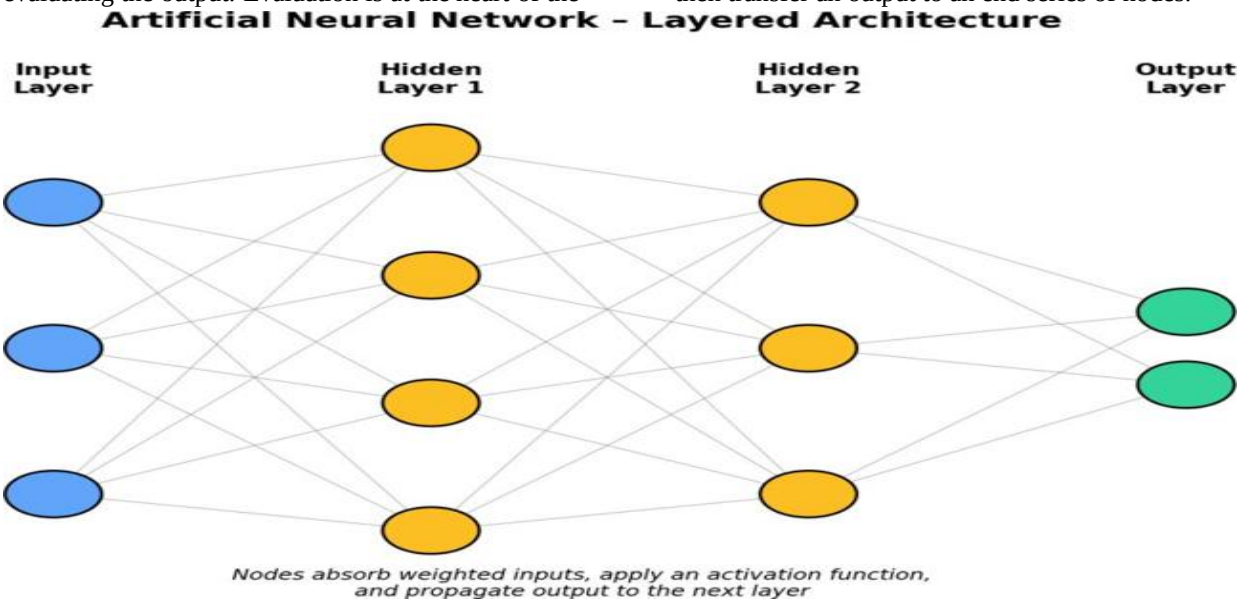


*Figure 6. Layered architecture of an artificial neural network: input, hidden, and output layers.*

## V. CHALLENGES AND ISSUES

A. Challenges

Efficient and effective data mining in large databases poses numerous requirements and great challenges to researchers and developers.

The issues involved include data mining methodology, user interaction, performance and scalability, and the processing of a large variety of data types.

Other issues include the exploration of data mining applications and their social impacts.

B. Issues

- Information poorness — The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. Data collected in large data repositories become “data tombs” — data archives that are seldom visited.
- Decision Making — Important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker's intuition. The decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data.
- Data Entry — Often systems rely on users or domain experts to manually input knowledge into knowledge bases. Unfortunately, this procedure is

prone to biases and errors, and is extremely time-consuming and costly.
- Bad Name — The term is actually a misnomer. Mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Data mining should have been more appropriately named "knowledge mining from data" which is unfortunately somewhat long.

## VI. RELATED WORK

A substantial body of prior work underpins the concepts and techniques revisited in this review. This section groups that work by theme — foundational texts, big data management, machine learning and deep learning surveys, and domain-specific applications — to situate the present paper's contribution within the existing literature.

### A. Foundational Data Mining Literature

The conceptual backbone of this review draws on established textbooks and survey papers that defined the field's core vocabulary and methodology. Han, Kamber, and Pei's text remains a standard reference for data mining concepts and techniques, while Witten and Frank's practical guide emphasizes applied machine learning tooling [1], [2]. Later survey efforts by Jain and Srivastava and by Kaur and Singh revisited these foundations specifically through the lens of big data, cataloguing technique families and comparing their suitability for large, heterogeneous datasets [4], [5].

### B. Big Data Management and Infrastructure

A second cluster of work addresses the infrastructural and governance dimensions of big data rather than the mining algorithms themselves. Van der Sloot, Broeders, and Schrijvers examine the legal and ethical boundaries that constrain large-scale data collection and use [3]. Wu, Zhu, Wu, and Ding's HACE-theorem framework characterizes big data through its heterogeneous, autonomous, complex, and evolving properties, offering a processing model for handling data at scale [10]. More recent engineering-oriented reviews, including Alaba et al. and the cloud-focused survey in the Qubahan Academic Journal, extend this discussion to real-life case studies and cloud-based mining architectures, highlighting scalability and system design as ongoing concerns [6], [9].

### C. Machine Learning and Deep Learning Surveys

Because the algorithmic techniques discussed in this paper (decision trees, evolutionary methods, and neural networks) are increasingly situated within broader machine learning and deep learning pipelines, several recent surveys inform this discussion. Batta Mahesh provides an accessible overview of classical machine learning algorithms [20], while Kumar et al. and Sharma et al. examine how machine learning techniques scale to big data settings and support predictive analytics across domains [7], [8]. On the deep learning side, Ahmad et al. and the Springer survey on deep learning fundamentals review architectural advances, and Li et al. focus specifically on active learning strategies for reducing labeling cost in large datasets; a complementary review of convolutional neural network architectures rounds out the deep learning perspective referenced in the neural network discussion above [16], [17], [18], [19].

### D. Domain-Specific Applications

Finally, a set of applied studies illustrates how the concepts reviewed here translate into specific domains. Moudi, Soleimani, and Hojjatinia survey deep learning-based intrusion detection systems for IoT data, directly relevant to the cybersecurity applications mentioned in this paper's abstract [11]. Ferhi et al. apply data mining classification techniques to diabetes diagnosis using the Orange data mining platform, an example of healthcare analytics in practice [12], while the bibliometric study of big data analytics in health research situates such work within a broader research landscape [15]. In the financial technology space, Yang, Ding, He, and Liu evaluate artificial-intelligence-based financial data mining techniques [13], and Abha and Handa apply data mining methods to assessing employment potential, illustrating the technique's reach beyond conventional commercial analytics [14].

Taken together, this literature shows a field that has moved from foundational, largely structured-data techniques toward heterogeneous, high-velocity, and increasingly deep-learning-driven pipelines — the trajectory this review aims to summarize and contextualize.

## VII. CONCLUSION

With very rapid data growth, research in the field of

big data and data mining is still growing rapidly too, those who struggling in big data will always find an increasement of complexity from big data. Therefore we conclude that the approach of data mining algorithms can still be improved. With the many problems that still exist now and issues that occur the possibility is still widely open.

REFERENCES


[1] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", ISBN 978-0-12-381479-1, 225 Wyman Street, Waltham, MA 02451, US, 2012.

[2] Ian H. Witten, Eibe Frank, "Data Mining Practical Machine Learning Tools and Techniques", 500 Sansome Street, Suite 400, San Francisco, CA 9411, 2015.

[3] Bart van der Sloot, Dennis Broeders, Erik Schrijvers, "Exploring the Boundaries of Big Data", Amsterdam, 2016.

[4] Nikita Jain, Vishal Srivastava, "Data Mining Techniques: A Survey Paper", eISSN: 2319-1163 | pISSN: 2321-7308, Rajasthan, India, 2013.

[5] Nirmal Kaur, Gurpinder Singh, "A Review Paper On Data Mining And Big Data", ISSN No. 0976-5697, Jalandhar, Punjab, India, 2017.

[6] F. Alaba et al., "A State-of-the-Art Review in Big Data Management Engineering: Real-Life Case Studies, Challenges, and Future Research Directions", Eng, 5(3), 1266-1297, 2024.

[7] R. Kumar et al., "Unlocking the Power of Machine Learning in Big Data: A Scoping Survey", ScienceDirect (Machine Learning with Applications), 2025.

[8] S. Sharma et al., "Big Data and Predictive Analytics: A Systematic Review of Applications", Artificial Intelligence Review, Springer, 2024.

[9] Comprehensive Survey of Big Data Mining Approaches in Cloud Systems, Qubahan Academic Journal, 2021.

[10] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data and its Techniques" (HACE theorem and Big Data processing model), 2022 (revised edition).

[11] M. Moudi, A. Soleimani, A. Hojjatinia, "A Survey of Intrusion Detection Systems Based on Deep Learning for IoT Data", November 2024.

[12] W. Ferhi, M. Hadjila, D. Moussaoui Djillali, A. B. Bouidaine, "Machine Learning-based Classification of Diabetes Disease: A Case Study with Orange Data Mining", Conference Paper, November 2023.

[13] C. Yang, D. Ding, J. He, W. Liu, "Application Evaluation of Financial Data Mining Technology Based on Artificial Intelligence Background", Conference Paper, October 2023.

[14] Abha, D. Handa, "Review of the Use of Data Mining Techniques for Assessing Employment Potential as Part of the Intelligent Computing", Conference Paper, January 2023.

[15] "Global Trends of Big Data Analytics in Health Research: A Bibliometric Study", PMC / BMC, 2024-2025.

[16] I. Ahmad et al., "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications", MDPI Information, 15(12):755, November 2024.

[17] "A Survey on Deep Learning Fundamentals", Artificial Intelligence Review, Springer, 2025.

[18] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, M. Okumura, "A Survey on Deep Active Learning: Recent Advances and New Frontiers", arXiv:2405.00334, IEEE TNNLS, 2024.

[19] "A Comprehensive Review of Convolutional Neural Networks: Foundations, Enhancements and Applications", Neural Computing and Applications, Springer, 2025.

[20] Batta Mahesh, "Machine Learning Algorithms - A Review", International Journal of Science and Research (IJSR), 9(1):381-386, 2020.