Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 06 Issue 01
2023

Research Article
https://doi.org/10.56979/601/2023

Data Mining Methods and Obstacles: A Comprehensive Analysis

Sunila Fatima Ahmad[1], Asad Hussain[2]*, Muhammad Asgher Nadeem[3], Arsalan Malik[4], Zubair Mushtaq[5], Kashif Hassan Khan[6], and Zahra Abbas[1]

[1]Institute of Information Technology, Quaid-e-Azam University, Islamabad, Pakistan.
[2]Department of Computer Science, University of Science and Technology, Bannu, 28100, Pakistan.
[3]Department of CS&IT, Thal University, Bhakkar, 30000, Pakistan.
[4]Department of Computer Science, COMSATS University Islamabad, Pakistan.
[5]Department of Computer Science and Information Technology, KTH University, Sweden.
[6]Department of Computer Science, University of Agriculture Faisalabad, Pakistan.
*Corresponding Author: Asad Hussain. Email: chasad1303@gmail.com

Received: August 22, 2023 **Accepted:** November 18, 2023 **Published:** December 05, 2023

Abstract: An in-depth analysis of the challenges and developments in the data mining industry is the aim of the study. To locate and assess relevant information on data mining approaches and difficulties, this study combines guided literature searches with real-world case studies. The research design includes information on the search terms, data sets, and selection standards for choosing which articles should be included and which should be omitted. Also, the study includes data extraction, outcomes interpretation, and paper quality assessment. The research focuses on data mining detection techniques and commercial challenges. The various data mining methods are discussed along with the challenges they face. Results from surveys, books, articles, published papers, finished projects, and reviews of earlier research are included in this analysis. According to the study's findings, data mining is a difficult occupation that is perpetually in need of advancement and creativity.

Keywords: Data Mining; Detection Methods; Obstacles; Data Analysis; Data Quality.

1. Introduction

Searching through large data sets for patterns and connections that may help in solving business problems via data analysis is known as data mining [1]. Businesses may anticipate future trends using data mining techniques and technology to make more informed business decisions. Data science's fundamental subject, data mining, uses complex analytical techniques to find valuable information in data sets.

Data mining is a step in the knowledge discovery in databases (KDD) procedure, a data science method for gathering, processing, and conducting more in-depth analyses of data. While they frequently function simultaneously, data gathering, and KDD are typically seen as distinct concepts. The act of detecting or discovering something is known as detection, while a method is a particular approach to carrying out an activity, usually one that calls for practical expertise. So, it describes a realistic way of discovering and detecting something. A technique is a method of performing a certain activity, such as the creation or execution of an art piece or any scientific technique. Technique refers to talent or competence in a certain sector, as well as an effective process of performing or attaining anything. A challenge is a serious topic of discussion or debate [2, 3].

We have thoroughly analyzed all our queries using data from these four carefully chosen databases, namely IEEE, ACM, Google Scholar, and Science Direct, which include research papers, publications, comprehensive surveys, and Articles from all years of research. The introduction of this survey is included in the first section. The Related View of data mining is covered in Section 2. The Methodology is described in Section 3 and is briefly explained in detail in various stages. The Conclusion is described in Section 4. The graphical representation of all the sections is represented in Figure 1.

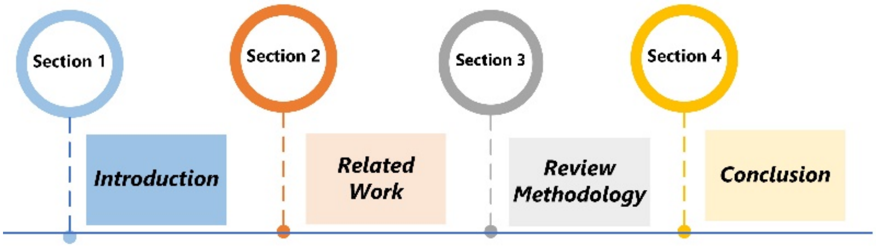

Figure 1. Structure of Analysis

2. Related Work

At the initial conference on the subject (KDD-1989), Gregory Piatetsky-Shapiro coined the name "knowledge discovery in databases," which the fields of AI and machine learning rapidly embraced. Nonetheless, the term "data mining" became increasingly common in the corporate and media spheres. Data mining sometimes referred to as "Big Data," is the computer-aided investigation and identification of patterns in enormous data sets [1].

It is a computing field that incorporates machine learning techniques, database theory, statistics, and data science. Data mining is commonplace, but its history goes back far further than Moneyball and Edward Snowden. Below are several landmarks and watershed moments in the evolution and integration of data mining with data science and large datasets.

Data mining sometimes referred to as "Big Data," is the computer-aided investigation and identification of patterns in enormous data set [4]. It is a computing field that incorporates machine learning techniques, database theory, statistics, and data science. The posthumous publication of Thomas Bayes' work on the Bayes' theorem, which connects current probabilities to prior probability, occurred in 1763. Since it makes it possible to understand and estimate probabilities in complex situations, it is crucial to data analysis and possibility [5].

Regression is a technique used by Carl Friedrich Gauss and Adrien-Marie Legendre to forecast celestial body trajectories in 1805. (Planets and comets) The purpose of regression analysis, which was used in this case using the least squares method, is to assess the relationships between variables. Among the most significant data mining methods is regression. The ability to collect and handle enormous amounts of data was made possible with the introduction of computers in 1936. On Computable Numbers, Alan Turing proposed the idea of a Universal Machine, a device that could do computations comparable to those of modern computers. The concepts created by Alan Turing are the foundation of the modern computer [6].

In 1943, Walter Pitts and Warren McCulloch created the first theoretical model of a neural network. The idea of a neuron within a network in a study titled "a logical calculus of the notions inherent in neurological function". Each of these neurons can take in information, process it, and then produce an output [7]. In 1965, Lawrence J. Fogel founded Decision Science, Inc. to address needs in the field of dynamic algorithms. It was the first company to apply evolutionary computing specifically to address problems in the real world. Terabytes and even petabytes of information could be stored and accessed in the 1970s, thanks to sophisticated database management systems. Data warehouses provide users with the ability to change their viewpoint when analyzing data from a commercial to an analytical one. Complex insights, however, are challenging to glean from massive multidimensional data warehouses [8].

The ground-breaking book on genetic algorithms, Resilience in Nature, and Human Ecosystems by John Henry Holland was published in 1975. The theoretical underpinnings and implications of this field of study are covered in this book for the first time. In the 1980s, HNC trademarked the term "database mining". The Data Mining Workspace technology was to be protected by the trademark. It was an out-of-date general tool for building neural network models. At this point, sophisticated algorithms can "learn" correlations from data, allowing specialists in the field to deduce the meaning of the links[9].

Knowledge Discovery in Databases was first used by Gregory Piatetsky-Shapiro in 1989. Around this time, he also helps co-found the first KDD workshop. The database industry adopted the phrase "data mining" in the 1990s. Retail businesses and the financial sector use data analysis to evaluate data and discover patterns to predict changes in borrowing rates, stock markets, increasing consumers and to grow their client base. To facilitate the development of nonlinear classifiers, the initial support vector machine was enhanced by Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik in 1992 [3]. Support vector computers are a supervised learning method that studies data and finds patterns for regression and classification. Knowledge Development Nuggets is a periodical started in 1993 by Gregory Piatetsky-Shapiro. K. D. Nuggets Its primary goal was to connect academics who took part in the KDD conference. Yet it seems like KDnuggets.com is suddenly seeing far more traffic.

While the term "data science" has been around since the 1960s, it was not until William S. Cleveland proposed it as a separate field of study in 2001. According to Create Data Science Teams, DJ Patil and Jeff Hammer Becher later modified the term to describe their roles at LinkedIn and Facebook. Money Ball, a 2003 book by Michael Lewis, changed the way that certain baseball league front offices conducted business. The Oakland Athletics utilized a data-driven, analytical approach to find underpaid, cheap players who have desired qualities. With barely a third of the payroll allocation, they skillfully built a team that led them to the semifinals in 2002 and 2003 [10, 11].

In February 2015, DJ Patil was appointed as the White House's first Chief Data Scientist. Data mining is widely used in many industries today, like marketing, research, engineering, and medicine, to name a few. Transactions with credit cards, stock market fluctuations, public safety, genomic decoding, and drug testing are just a few examples of data mining uses. Phrases like "Big Data" are now widely used due to the spread of data acquisition equipment and the decrease in data collection costs. One of the methods now in use that is being actively investigated is deep learning. It is sparking some of the biggest challenges in data analysis, computer and data science, and machine intelligence since it is far better than previous techniques at capturing connections and complex patterns [12].

3. Methodology

The purpose of this research was to conduct a comprehensive survey with a sharp focus on data mining. Figure 2 illustrates our Survey Review Methodology Existing PRISMA criteria serve as the foundation for this investigation [13].

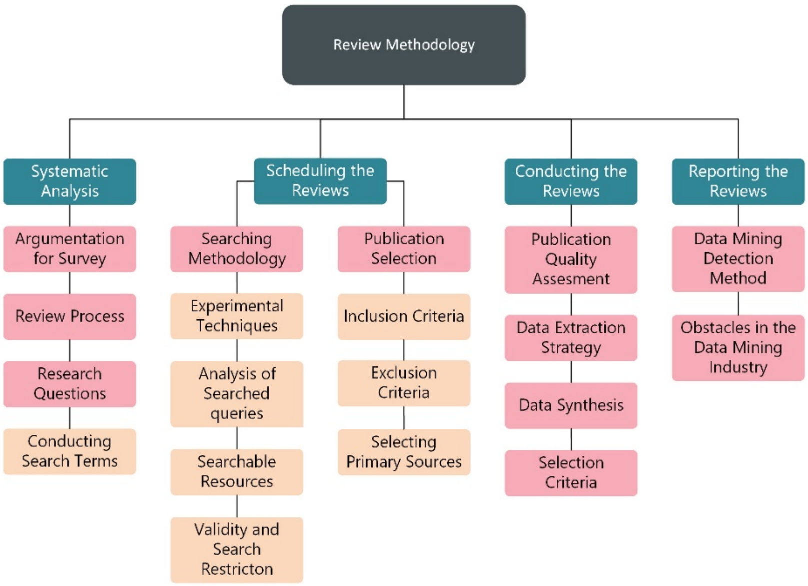

Figure 2. Proposed Methodology

3.1 Systematic Analysis

3.1.1 Argumentation for Survey

Historically, data mining detection strategies and obstacles in data mining are examined in depth in several scholarly works. In this survey, we provide a comprehensive Systematic literature analysis of detection strategies and problems in data mining. These questions pertain to these crucial concerns. The first question concerns data mining detection strategies. The second question relates to data mining challenges.

3.1.2 Review Process

The review approach of this research is based on the well-known PRISMA criteria [13]. The creation of this extensive literature analysis is mostly based on suggestions from other research[14]. We looked at surveys, publications, journals, research papers, and other people's studies to get more details.

3.1.3 Research Questions

The following research issues prompted the work presented in this research paper:

- What are the detection techniques of data mining?
- What are the challenges in the field of data mining?

3.1.4 Conducting Search Terms

This information will help us produce a good search keyword for our research questions.

Population: Data Mining

Intervention: Detection Techniques

Outcomes of relevance: Challenges in the field of data mining

3.2 Scheduling the Reviews

3.2.1 Searching Methodology

3.2.1.1 Experimental Techniques

The following search term was used in a sample search on all the digital libraries.

- Data Mining: ("Detection" OR "Noticing") AND ("Techniques" OR "Approach" OR "Method") AND ("Data mining" OR "Data collection").
- Obstacles: (“Challenges” OR “Objection” OR “Obstacles”) AND (“Sphere” OR “Range” OR “Plot”) AND ("Data mining" OR "Data collection").

The articles that were located using these search terms will be used to assist in designing and validating the key search terms.

3.2.1.2 Analysis of Searched queries

Search terms are generated using the following method:

- Parse the Research Issue to find key terms that describe the sample, treatment, and outcome.
- It is recommended to use the Research Questions to find alternative orthography and alternatives for relevant phrases.
- Review the paper's keywords.
- If the database allows for it, use the OR technique to combine different forms of the same word, along with the AND function to merge various parts of a keyword.

3.2.1.3 Searchable Resources

The following online resources and databases are reviewed.

- IEEE Explore
- Science Direct
- ACM Digital Library
- Google Scholar

3.2.1.4 Validity and Search Restriction

The following query was used to do a representative search on all online libraries.

- ("Detection" OR "Noticing") AND ("Techniques" OR "Approach" OR "Method") AND ("Data mining" OR "Data collection").
- ("Challenges" OR "Objection" OR "Obstacles") AND ("Sphere" OR "Range" OR "Plot") AND ("Data mining" OR "Data collection").

The papers that were located using these search term
will be used to assist in designing and validating the key search terms.

3.2.2 Publication Selection Search

Publications will be chosen using a combination of primary source selection, inclusion criteria, and exclusion criteria. Our primary objective in performing this selection technique on manuscripts is to narrow the results of our query to just those papers that are of relevance to our research. Papers, discussions, and publications not relevant to Data Mining will be disregarded. Books, papers, and articles that do not deal with Data Mining will not be considered.

3.2.3 Inclusion Criteria

Just the articles, reports, and books that are returned in a search are chosen for inclusion. Studies on the use of AR in research alone will be considered.

Requirements for inclusion consist of:

- The documentation of methods for detecting data mining attacks.
- Research outlining the challenges in data mining.

3.2.4 Exclusion Criteria

The publications (essays, reports, and novels) that will not be examined are chosen using exclusion criteria based on the search term. The following prerequisites must be fulfilled:

- Research that does not address research questions.
- Research that ignores the detection techniques and challenges faced by the data mining sector.
- Research that does not use data mining.

3.2.4.1 Selecting Primary Sources

To pick the primary sources first, the headings, keyword phrases, and descriptions of the information that has been retrieved will be employed. Any content that is not relevant to the study topics will be rejected or removed from this evaluation. The integration parameters will be implemented to the original data selected during the initial screening step by examining the whole transcripts of published articles.

The additional assessor will be presented with the case if the inclusion/exclusion decision is unclear in any way. The third assessor will evaluate the method. Each primary source is required to keep precise records of the items it includes and excludes. In this part, we will discuss the reasoning for including or removing the major contributor from the overall rating.

3.3 Conducting the Reviews

3.3.1 Publication Quality Assessment

The quality assessment of selected publications follows the final selection. At about the same time that data is being extracted, it is being evaluated.

The following responses will be included on the quality evaluation checklist as "Yes", "No," "Partial", or "NA":

- Is there a proven technique for identifying data mining attacks?
- Has it been identified as an issue in data mining?

3.3.2 Data Extraction Strategy

This survey's main goal is to collect information from academic articles that will help answer our research questions.

To do our analysis, we will need the following data from each research article:

- Details about the article (title, authors, name of the journal or conference, and other pertinent information).
- Data that addresses our study questions.

The following data will be obtained to respond to our study questions.

- RQ#1: Historical context and approaches for data

data mining detection.

- RQ#2: Contextual Information and Data Mining Difficulties.

3.3.3 Data Synthesis

There will be two phases to the data synthesis: Q1 and Q2. In this first section, we will talk about several techniques for spotting anomalies in data mining. For the second part, we will stick with the data mining concerns raised in the second question.

3.3.4 Selection Criteria

Four steps of selection criteria are shown in Figure 3.

3.3.4.1 First phase

To create the search phrases, we consulted several literature studies on data mining methods and difficulties. Reviewing the next part came first. The first phase's collection of papers totals 4,506.

3.3.4.2 Key Terms

The various repositories are combed through using the following search phrases and queries for each query.

- Data Mining: (“Data mining” OR “Data retrieval” OR “Data analytics” OR “Data collection” OR “Data acquisition” OR “Data capturing” OR “Data accumulation” OR “Data compilation” OR “Data gathering” OR “Fact-finding” OR “Data harvesting” OR “Data recording”)
- Methods: (“Method” OR “Approach” OR “Procedure” OR “Process” OR “System” OR “Method of working” OR “Tactics” OR “Strategy” OR “Practice” OR “Plan” OR “Manner”)
- Obstacles: (“Objection” OR “Protest” OR “Test” OR “Threat” OR “Claiming” OR “Dare” OR “Interrogation” OR “Provocation” OR “Demanding” OR “Defiance”)
- Field: (“Range” OR “Sphere” OR “Plot” OR “Enclosure” OR “Track” OR “Court” OR “Course” OR “Arena” OR “Stadium” OR “Park” OR “Playground”)

3.3.4.3 Second phase

Due to the massive amount of research papers collected in the first stage, the second stage comprises abstract-based filtering. Because of the filtering options provided by the reference abstract, we were able to identify studies that were published in many locations and in different iterations. Sixty-five entries from multiple repositories were the previous cap for this round.

3.3.4.4 Third phase

At this stage "abstract" were given priority over detailed metadata. It is easy to exclude extraneous information by reading the abstract of each publication; using this technique, 10 publications were eliminated, and the remaining 55 research materials were in our repository.

3.3.4.5 Fourth phase

We then went back and read over the remaining articles. Publications that did not make the quality cut or were beyond the scope of our study were omitted using this procedure. At this stage, 17 studies were taken out of the study, leaving 41 research papers. The method for selecting the survey is provided

3.3.4.6 Reporting the Reviews

The outcomes of our exhaustive, in-depth research were noteworthy. This part consists of a summary of our results, followed by a discussion. Regarding the articles' conclusions, all the research questions were used for analysis. Table 1 provide a few statistical details before outlining the study's concerns. We fully explain the analytical outcomes before posing open questions to researchers in the field.

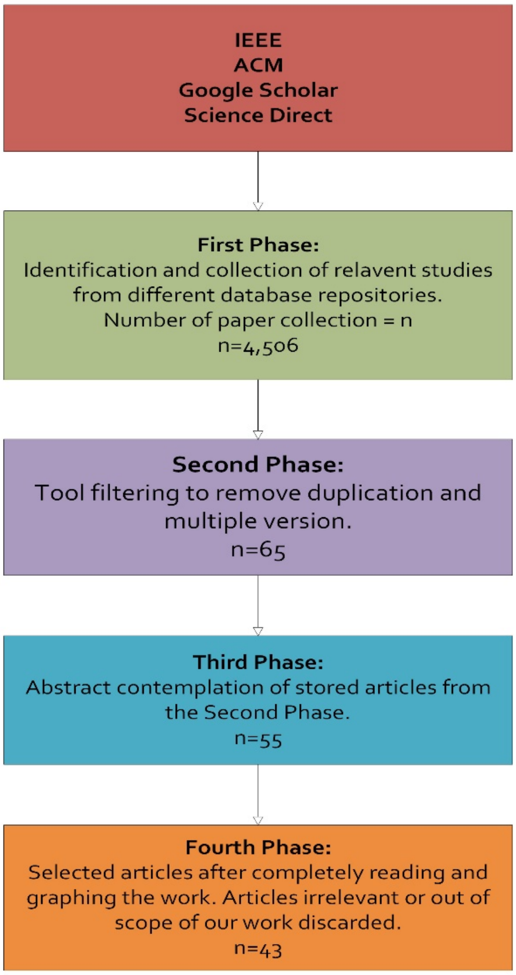
IEEE
ACM
Google Scholar
Science Direct
First Phase:
Identification and collection of relavent studies from different database repositories.
Number of paper collection = n
n=4,506
Second Phase:
Tool filtering to remove duplication and multiple version.
n=65
Third Phase:
Abstract contemplation of stored articles from the Second Phase.
n=55
Fourth Phase:
Selected articles after completely reading and graphing the work. Articles irrelevant or out of scope of our work discarded.
n=43

Figure 3. Selection procedure with phase-by-phase filtering**.**

Table 1. Sources of Information

Sr.#	Questions	ACM	IEEE	Google Scholar	Science Direct	Total Papers
1	Q1	8	10	13	2	33
2	Q2	2	3	5	0	10

3.4 Data Mining Detection Methods

It is necessary to clean and prepare data before mining it. To be utilized by different analytic approaches, raw data must be organized and cleansed. Most Popular Data Mining Techniques are listed in Table 2. Data modeling, transformation, immigration, Extract, Transform, and Load (ETL), data aggregation, and consolidation are just a few of the processes involved in data cleaning and preparation. To determine data's best usage, it is essential to understand its basic properties and facets. The value of data preparation and cleaning is clear from a business standpoint [1].

A company's operations are either pointless or of inferior quality and unstable if this initial stage is skipped. Businesses need to trust their information and the decisions they make consequently. These processes are also necessary for appropriate data governance and data quality [5, 9].

A key data mining technology is pattern recognition. It requires identifying and keeping an eye on patterns and trends in data so that judgments about business outcomes may be made with knowledge. For example, it makes sense to capitalize on information whenever a corporation notices a trend in sales data [15, 16]. If a business discovers a specific product that appeals to a particular demographic more than others, it may use this knowledge to produce related products or services or to simply keep additional supplies of the original item on hand.

Many characteristics connected to various data types must be evaluated to use classification data mining techniques [4]. Organizations may classify or categorize related data after determining the key characteristics of various data types. This is necessary to recognize, for example, sensitive data that businesses may wish to protect or exclude from papers.

One technique for data mining that is statistically based is an association [17]. It displays the relationship between certain data (or data-driven events) and some other information or data-driven occurrences. It is analogous to the machine learning theory of co-occurrence, where the presence of one statistical event implies the likelihood of another. The concepts of association and correlation in statistics are similar. This demonstrates how data analysis may be used to discover a connection between two informational events, such as the fact that purchasing sandwiches are often matched by purchasing French fries.

Any anomalies in datasets are found via outlier identification [18]. When businesses produce entropy in their information, it is easier to understand why they happen and to prepare for them in the future, which helps them achieve their objectives more successfully. For instance, businesses may use this statistic to maximize their purchases for the rest of the day if there is a spike in the usage of database technology for credit card payments at a certain time [19]. Clustering is a kind of analytical tool that relies on ways to understand data visually. Clustering approaches make use of graphics to show where information dispersion is relative to diverse types of metrics. Clustering methods commonly depict the information's dispersion in a range of hues. For cluster analytics, graph approaches should be used. Users may discover patterns relevant to their current business objectives by using diagrams and clusters to understand how data is distributed.

Finding out the kind of variable associations in a dataset may be done with the use of regression algorithms [6]. These relationships could be causal in certain circumstances but just correlative in others. Regression is a fundamental white-box method for revealing how variables are related to one another. Regression algorithms are among the prediction and data modeling components. Prediction is one of analytics' four main types, which is also an amazingly effective portion of data mining. Predictive analytics considers future trends that have been observed in current or historical data. Because of this, it provides companies with a glimpse into probable future data patterns. There are several applications for predictive analytics. Many innovative technologies today make use of AI and ML. Predictive analytics may, however, be carried out using simpler algorithms without the requirement for these methods.

The goal of this data mining strategy is to identify a chain of events that happens consecutively. Particularly for transactional data mining, it is advantageous. This method may reveal, for example, if a customer buys a pair of shoes, what other articles of apparel are they likely to make subsequent purchases? If customers understand sequential patterns, businesses may suggest additional goods to them to enhance sales [9, 20].

A decision tree algorithm is a kind of projection model that helps companies gather data effectively. A decision tree is a part of machine learning, but owing to its extreme simplicity, it is more frequently described as a white-box machine learning approach. A decision tree may help customers see how several options affect the result. When many decision tree models are combined, a random forest is created, which may be used as a kind of predictive analytics.

Table 2. Data Mining Detection Techniques

Sr.#	**Papers**	**Data Mining Detection Techniques**
1	[1, 9, 21, 22]	Data cleaning and preparation
2	[15, 16, 23]	Tracking
3	[4]	Classification
4	[17]	Association
5	[18]	Outlier
6	[19]	Clustering
7	[6]	Regression
8	[6, 24]	Prediction
9	[6, 9, 20, 24]	Sequential Patterns
10	[8]	Decision Trees
11	[2]	Statistical Techniques
12	[25]	Visualizations
13	[7, 26]	Neural Networks

Black box methods of ML refer to sophisticated random forest models because it might be difficult to understand their findings in the context of their inputs. This basic kind of ensemble modeling is often more accurate than using just decision trees [8].

Most analyses used in the procedure of data mining depend on statistical methods [2]. Numerous analytical models are based on statistical ideas that provide numerical values appropriate for certain business objectives. Neural networks, for example, use complex statistics based on several weights and metrics to determine whether a picture represents a dog or a cat in image recognition systems. One of artificial intelligence's two main specialties is the research of predictive methods. Although some statistical methods utilize static models, others that use ML become better with time. An additional crucial element of data mining is data visualization. They provide customers with access to information based on data from observable sensory experiences. Modern data visualizations are dynamic, appropriate for actual data streaming, and characterized by a range of colors that denote unique data patterns and trends.

The best way to use data visualizations to get data mining insights is via dashboards [25]. Organizations may create dashboards based on a variety of indicators and utilize graphics to demonstrate data patterns rather than only relying on the quantitative findings of statistical models. A neural network is a kind of algorithm for machine learning that is often used with deep learning and artificial intelligence. The most dependable machine learning models currently being used are neural networks. Its title is derived from the fact that the different layers of these structures mimic the way that cells in the human brain work. While neural networks may be a beneficial tool for data mining, companies should utilize them with caution. It might be challenging to understand how a neural network created an output since some neural network models are quite complex [7].

3.4.1 Obstacles in the Data Mining Industry

Data collection-sharing is used in dynamic approaches; therefore, it takes a lot of security to always keep sensitive and confidential data safe. Data mining is a method for extracting information from enormous amounts of data. This knowledge of the existing world is loud, inconclusive, and varied. Large volumes of often erroneous or faulty data will be present. These problems may be the result of either technical or human error. Obstacles in the field of Data Mining are shown in Table 3. Real data is not kept in a specific location; it may be found on several systems, the internet, or even databases [10].

The main organizational and technological challenges in moving all the information to a consolidated data repository are these [7]. Correct data is stored throughout several phases under the situation of dispersed processing. Real data is diverse and may comprise time series, geographical data, temporal data, complicated data, audio or video, photos, natural language text, and more. To isolate critical information, new tools and methods would often need to be developed. The effectiveness of the used methods and algorithms will determine how the data-gathering architecture is presented.

Table 3. Obstacles in the Data Mining Industry

Sr.#	**Papers**	**Obstacles in the Data Mining Industry**
1	[10, 27]	Security and Social Challenges
2	[7, 28]	Noisy and Incomplete Data
3	[7, 28-30]	Distributed Data
4	[31, 32]	Complex Data
5	[33, 34]	Performance
6	[11, 35]	Scalability and Efficiency of the Algorithms
7	[36]	Improvement of Mining Algorithms
8	[37, 38]	Incorporation of Background Knowledge
9	[25, 39-41]	Data Visualization
10	[42, 43]	Mining Dependent on Level of Abstraction

If the methodologies or algorithms used are insufficient, it will have a negative impact on how the measure is presented. The data mining approach must be both scalable and successful if it is to be used to extract useful information from the vast amounts of data included in the dataset. Factors such as the intricacy of data mining methods, the quantity of the dataset, and the total data flow stimulate the spread and innovation of contemporaneous solutions for data mining. The process of gathering and incorporating foundational information is unexpected [30, 33].

Predictive activities may produce more accurate projections, while explanatory projects may include more insightful findings. Including historical context in data mining projects may help uncover more reliable and accurate results. Data visualization is a crucial phase in the data mining process because it is the first point of contact with the client and must make a good impression. To the end-user, however, it might be difficult to provide the necessary information in a straightforward and simple manner. It is important to use data perception techniques since the quantity of output and input data are both phenomenally successful and complicated. For both organizations and individuals, data mining may result in serious problems with management, confidentiality, and data security [3].

If the client considers the material fascinating or more reasonable in general, then the data mining tool will determine that it is valuable. A good translation of data representations may assist with data mining results and is essential for gaining a thorough understanding of the needs inherent in any data mining activity [36]. To get a fantastic viewpoint, several studies are undertaken on huge data sets that evolve and display mined information. Strategies for collecting information should be community-oriented, as this allows users to focus on activities like optimizing, presenting, and pattern-finding based on returning discoveries. Background research may be utilized to characterize recognized patterns and direct the research operations [37]. The management and treatment of data noise as well as the domain's dimensionality are two issues that might make it challenging to analyze vast volumes of data. They include the variety of data at hand and the adaptability of the mining technique, among other things.

4. Conclusion

Despite the large amount of study that has been carried out on the topic of data mining, there has yet to be enough information available in the form of an all-encompassing overview that discusses the several approaches and challenges that are specific to this industry. The results of our comprehensive analysis of the relevant literature indicate that most of the studies provide answers that are comparable, underscoring the want for a greater variety of research methods. To fill this void, we used a qualitative methodology to give a comprehensive and in-depth overview of the various data mining approaches as well as the challenges that come along with them. This article presents a complete review of the issues that are present in the industry of data mining as well as the numerous ways that may be employed to solve those challenges. We believe that by giving this in-depth study, we will be able to help foster the development of understanding in this crucial sector and make it easier for others to do research in the field of data mining in the future.

References

1. Jain, N. and V. Srivastava, Data mining techniques: a survey paper. IJRET: International Journal of Research in Engineering and Technology, 2013. 2(11): p. 2319-1163.
2. Pujari, A.K., Data mining techniques. 2001: Universities press.
3. Purwar, A. and S.K. Singh. Issues in data mining: A comprehensive survey. in 2014 IEEE International Conference on Computational Intelligence and Computing Research. 2014. IEEE.
4. Suguitan, A.S. and L.N. Dacaymat. Vehicle Image Classification Using Data Mining Techniques. in Proceedings of the 2nd International Conference on Computer Science and Software Engineering. 2019.
5. Hegland, M., Data mining techniques. Acta numerica, 2001. 10: p. 313-355.
6. Wu, C.-s.M., M. Badshah, and V. Bhagwat. Heart disease prediction using data mining techniques. in Proceedings of the 2019 2nd international conference on data science and information technology. 2019.
7. Cantoni, V., L. Lombardi, and P. Lombardi. Challenges for data mining in distributed sensor networks. in 18th International Conference on Pattern Recognition (ICPR'06). 2006. IEEE.
8. Jain, N., et al. Computerized forensic approach using data mining techniques. in Proceedings of the ACM Symposium on Women in Research 2016. 2016.
9. Han, J., J. Pei, and H. Tong, Data mining: concepts and techniques. 2022: Morgan kaufmann.
10. Kuhlmann, M., D. Shohat, and G. Schimpf. Role mining-revealing business roles for security administration using data mining technology. in Proceedings of the eighth ACM symposium on Access control models and technologies. 2003.
11. Rygielski, C., J.-C. Wang, and D.C. Yen, Data mining techniques for customer relationship management. Technology in society, 2002. 24(4): p. 483-502.
12. Elgendy, N. and A. Elragal. Big data analytics: a literature review paper. in Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings 14. 2014. Springer.
13. Kitchenham, B., Procedures for performing systematic reviews. Keele, UK, Keele University, 2004. 33(2004): p. 1-26.
14. Hussain, A., et al., Computer Malware Classification, Factors, and Detection Techniques: A Systematic Literature Review (SLR). International Journal of Innovations in Science & Technology, 2022. 4(3): p. 899-918.
15. Dias, L.F.C. and F.S. Parreiras. Comparing data mining techniques for anti-money laundering. in Proceedings of the XV Brazilian Symposium on Information Systems. 2019.
16. Zaki, M.J. and L. Wong, Data mining techniques, in Selected Topics in Post-Genome Knowledge Discovery. 2004, World Scientific. p. 125-163.
17. Liu, X., Glaucoma screening using data mining techniques. ACM SIGBIO Newsletter, 1998. 18(3): p. 7-7.
18. Olson, D.L. and D. Delen, Advanced data mining techniques. 2008: Springer Science & Business Media.
19. Berkhin, P., A survey of clustering data mining techniques, in Grouping multidimensional data: Recent advances in clustering. 2006, Springer. p. 25-71.
20. Obenshain, M.K., Application of data mining techniques to healthcare data. Infection Control & Hospital Epidemiology, 2004. 25(8): p. 690-695.
21. Al Fanah, M. and M.A. Ansari. Understanding E-learners' Behaviour Using Data Mining Techniques. in Proceedings of the 2019 International Conference on Big Data and Education. 2019.
22. Alsharif, A.H. and N. Philip. Data mining technique for the enhanced smoking cessation management system (Smoke Mind). in 2016 International Conference on Engineering & MIS (ICEMIS). 2016. IEEE.
23. Ye, Y., et al., A survey on malware detection using data mining techniques. ACM Computing Surveys (CSUR), 2017. 50(3): p. 1-40.
24. Berry, M.J. and G.S. Linoff, Data mining techniques. 2009: John Wiley & Sons.
25. Jin, H. and H. Liu. Research on visualization techniques in data mining. in 2009 International Conference on Computational Intelligence and Software Engineering. 2009. IEEE.
26. Fan, W. and A. Bifet, Mining big data: current status, and forecast to the future. ACM SIGKDD explorations newsletter, 2013. 14(2): p. 1-5.
27. Ko, A.J. Mining the mind, minding the mine: grand challenges in comprehension and mining. in Proceedings of the 26th Conference on Program Comprehension. 2018.
28. Jun Lee, S. and K. Siau, A review of data mining techniques. Industrial Management & Data Systems, 2001. 101(1): p. 41-46.
29. Mucherino, A., P. Papajorgji, and P.M. Pardalos, A survey of data mining techniques applied to agriculture. Operational Research, 2009. 9: p. 121-140.
30. Sebastian, L.R., S. Babu, and J.J. Kizhakkethottam. Challenges with big data mining: A review. in 2015 International Conference on Soft-Computing and Networks Security (ICSNS). 2015. IEEE.
31. Raju, G., et al. Vulnerability assessment of machine learning based malware classification models. in Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2019.
32. Ruan, D., et al., Intelligent data mining: techniques and applications. Vol. 5. 2005: Springer Science & Business Media.
33. Brünink, M. and D.S. Rosenblum. Mining performance specifications. in Proceedings of the 2016 24th acm sigsoft international symposium on foundations of software engineering. 2016.

34. Singh, J. and J. Singh, A survey on machine learning-based malware detection in executable files. Journal of Systems Architecture, 2021. 112: p. 101861.
35. Shirahama, K., Intelligent video processing using data mining techniques. ACM SIGMultimedia Records, 2011. 3(2): p. 7-9.
36. Karagiannis, I. and M. Satratzemi. Comparing LMS and AEHS: challenges for improvement with exploitation of data mining. in 2014 IEEE 14th international conference on advanced learning technologies. 2014. IEEE.
37. Chen, F., et al., Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 2015. 11(8): p. 431047.
38. Liao, S.-H., P.-H. Chu, and P.-Y. Hsiao, Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications, 2012. 39(12): p. 11303-11311.
39. Ajibade, S.S. and A. Adediran, An overview of big data visualization techniques in data mining. International Journal of Computer Science and Information Technology Research, 2016. 4(3): p. 105-113.
40. Keim, D.A., Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics, 2002. 8(1): p. 1-8.
41. Martínez-Martínez, J.M., et al., A new visualization tool for data mining techniques. Progress in Artificial Intelligence, 2016. 5: p. 137-154.
42. Baier, T., J. Mendling, and M. Weske, Bridging abstraction layers in process mining. Information Systems, 2014. 46: p. 123-139.
43. Diba, K., et al., Extraction, correlation, and abstraction of event data for process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020. 10(3): p. e1346.