International Journal of Basic and Applied Sciences, 15 (1) (2026) 177-192

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/472vp602
Research paper

Software Concept and Knowledge Extraction for Data Mining In Healthcare Sectors

Hamid Jassam Mohammed *

Department of Cyber Security Engineering, College of Engineering, Al-Karkh University of Science, Baghdad, IRAQ
**Corresponding author E-mail: hamidj.mohammed@kus.edu.iq*

Received: October 13, 2025, Accepted: January 2, 2026, Published: January 26, 2026

Abstract

A technique called data mining makes it easy to extract valuable insights from large databases. It is the procedure of acquiring knowledge from information. This technical capability has generated much interest in society recently, especially in the information sector. Focusing on the healthcare industry, which has significant implementation issues, this article emphasizes the need for software in data mining by offering important backing for the development of effective and maintainable systems. Designing dependable systems, specifying both functional and non-functional needs, creating efficient algorithms, guaranteeing sustainable development, improving performance, and facilitating inte-gration are important elements of software pertinent to this field. The study underlines that software underpins the execution of good and efficient data mining techniques. Additionally, this research will examine uses of data mining in healthcare information as well as related problems. This descriptive study examines the ideas, problems, and techniques applied in data mining, particularly in the healthcare sector.

Keywords: *Software; Data Mining; Knowledge Management; Information Technology; Health Care.*

1. Introduction

Finding valuable insights from large datasets using data mining is the process known as data mining. Finding a balance between institutional events and the great amount of client data kept in databases requires [1].
particularly examines the healthcare industry, which generates huge amounts of administrative data including patient information, hospital specifics, bed costs, claims, and other information. As electronic medical records, computerised disease management, and clinical trial results become more readily accessible, clinical data is growing rapidly. For healthcare institutions, these data offer a strategic asset; data mining techniques are meant to explore them more thoroughly [2], [3].
These methods let experts explore hidden patterns and connections within data stores, which they may then research and employ to gain insights and steer healthcare choices. It should be stressed that decisions on healthcare should be made by healthcare professionals alone [4]; IT system specialists should not decide.
According to the American Medical Informatics Association, health informatics encompasses all aspects of knowing and promoting the organization, analysis, administration, and effective use of data in healthcare. Similarly, the Canadian Health Informatics Association regards it as the confluence of management, information technology (IM/IT), and clinical practice aiming to enhance health. Both definitions span a range of technologies, from the development of electronic health data systems to the installation of wireless networks in hospitals. A more exact definition from the National Library of Medicine is the study of information including the analysis and communication of medical data as well as computer applications to several aspects of healthcare and medicine. Health IT here should be understood as expressly concentrating on the analysis and distribution of medical information, excluding purely IT operations such network installation in hospitals. Zaiane offers an even more detailed definition, breaking health informatics into four subfields: “Health Informatics is the computerization of health information to support and optimize health services administration, clinical care, medical research, and training. It involves the application of computing and communication technologies to optimize health information processing through collection, storage, retrieval (in the right time and place), analysis, and decision support for administrators, clinicians, researchers, and educators in medicine.” [5].
The success of expert data mining systems relies not only on effective algorithms but also on the ability of users to leverage these systems to extract meaningful reports from the data [6].
This article is divided into two main sections: the first explains the terms related to data extraction, including their scientific definitions, tasks, and architecture. The second section focuses on data mining management applications, providing detailed examples from the healthcare sector.

2. Data Mining

Among the important phases of knowledge discovery from databases (KDD) is data mining, which helps uncover meaningful patterns or trends from data. Although typically employed together, data mining and KDD describe different ideas. KDD is knowledge discovery in databases; it is a multi-step process in data mining aimed at extracting useful information from big datasets. Focus is on algorithms for identifying significant patterns from the enormous amount of data in databases; data mining relates to discovering new patterns [7, 8]. It comprises several stages aimed at transforming raw data into actionable insights or patterns, as shown in Figure 1. Here is a breakdown of the usual stages in the KDD process:

- Data selection is the process whereby the appropriate data from the existing database is chosen. The goal is to concentrate on information that will help to answer the particular challenge or query being addressed [9].
- Data often contains noise, missing values, and variations; pre-processing is consequently, cleaning the data and resolving these issues call for preprocessing. Eliminating duplicates, imputing missing values, normalizing or scaling the data, and controlling categorical variables are among typical activities [10].
- Transformation: Data might be changed or merged at this stage to increase the mining process's effectiveness. Usually employed to generate a better depiction for data mining [11] are techniques like aggregation, generalization, or normalization.
- Data Mining: This main step is when algorithms are run on preprocessed data to discover models or patterns. It involves a range of methods, including regression, anomaly detection, association rule mining, clustering, and classification. This entails selecting an acceptable data mining technique for data patterns or data pattern extraction [12].
- Interpretation/evaluation: This entails converting helpful models into words that people can comprehend [13] or translating information models to remove redundant or extraneous models.

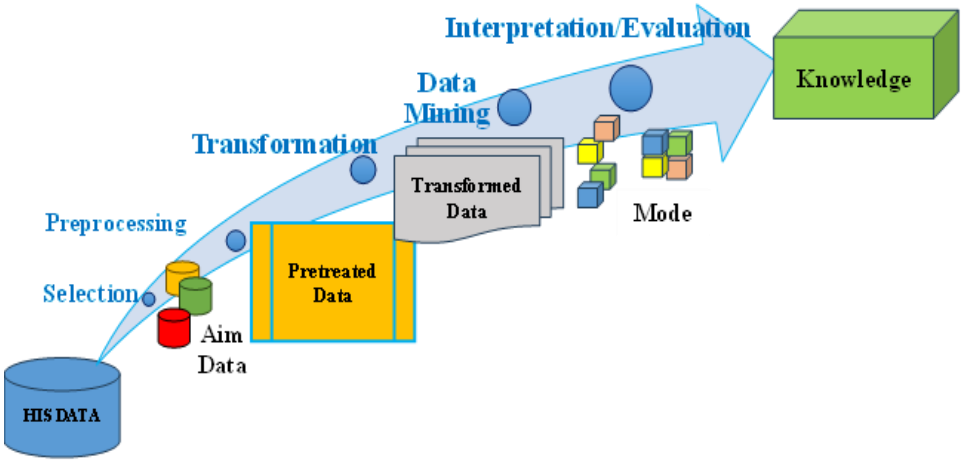

Fig. 1: Data Mining and the KDD Process.

2.1. Importance of data mining

Data mining involves retrieving information from extensive databases. The growing necessity for data has led to two primary motivations for employing data mining techniques:

- An excess of data accompanied by a scarcity of useful insights.
- The genuine requirement to derive meaningful information from the data and to analyze it effectively.

This scenario, characterized by an abundance of data yet a lack of actionable information, is often referred to as being "data-rich but information-poor." Consequently, the vast amounts of data stored in large repositories can become what are known as “data tombs,” essentially archives that are infrequently accessed [14].

2.2. History of data mining

The term data mining is used to denote a technical concept which first appeared in the 1990s, but the historical background of data mining has a long history. There are three main domains on which data mining is based, which include classical statistics, artificial intelligence (AI) and machine learning. The brief description of these concepts is given below:

- Statistics is the basis of several methods used in data mining including regression analysis, normal distribution, standard deviation, variance, discriminant analysis, cluster analysis, and confidence intervals. These techniques are used in order to analyze the data and examine the connections between various data sets.
- Artificial Intelligence (AI) is based on heuristics and not statistical procedures and is intended to imitate the human process of thinking to solve statistical problems. Other high-tech business uses of artificial intelligence have been incorporated in various commercial applications, such as query optimization in relational database management systems.
- Machine Learning uses elements of both statistics and AI. It is seen as a development of AI since it involves the combination of heuristics and sophisticated statistical methods. Machine learning aims to provide the opportunity to make computer programs better comprehend the information passed through them so that the programs themselves can make reasonable choices depending on the features of the processed data by relying on the key principles of statistics and with the help of complex heuristics and sophisticated algorithms implemented by AI.

Simply, data mining is the use of machine learning techniques in business. It can be described as a combination of the past and present statistics, AI, and machine learning developments. Taken together, these techniques are combined to be used in analyzing data to identify new trends or patterns that have not been realized before [16], [17].

2.3. Data mining tasks

Tasks in data mining can be classified into different types that can be generalized into two main categories, namely descriptive and predictive. These activities play a significant role in identifying hidden trends, patterns, and relationships in datasets.

a) Descriptive Tasks

Descriptive tasks are aimed at the analysis and summarization of the dataset in order to find its patterns and relations without predictions.

- Classification - This is the process of grouping data to a given pre-existing category or group of data. One of this is the classification of emails as spam or non-spam.
- Clustering: This is a method that groups similar data points according to the common features but does not employ predefined labels. One of the typical ones is customer segmentation based on purchasing behavior.
- Association Rule Mining: This is a technique that reveals associations or patterns between items in large data sets. A famous example is the market basket analysis, which shows that bread buyers are also prone to buy butter.
- Summarization: This involves finding the general characteristics of a dataset, e.g. finding measurements like mean, median, or variance, which provide a general impression of the characteristics of the data.

b) Predictive Tasks

Predictive tasks make use of past to determine future values or unknown values.

- Regression: the method is used to predict a continuous dependent on the independent variables. As an example, it can be applied to project the real estate prices on the basis of size and location.
- Classification (predictive aspect): This is a mechanism that makes predictions based on past information and the prediction is based on the desired category where the observation belongs based on the workings of spam detectors but in this case it makes predictions based on previous data.

c) Anomaly Detection (Outlier Detection)

This process is used to detect atypical points of data that do not fit in the general pattern or trend of the data set. Its uses are specifically useful in such areas as fraud detection, network security, as well as monitoring the performance of the system.

d) Sequential Pattern Mining

The method reveals trends that are used in a particular order within a data set. An illustrative case would be finding sequences of events or purchases which often happen in combination over a period of time.

e) Text Mining

Text mining is a technique of examining unstructured textual information to remove important information. Such techniques as sentiment analysis and topic modeling are used to disclose the insights based on textual content.

f) Time Series Analysis

This is an analysis that involves looking at data that is taken through time to identify some trends, cycles or seasonal changes. It is commonly applied in the forecasting aspect such as prediction of stock prices or sales trends within a given time.

These functions can be used effectively to help organizations extract useful information about raw data, thus empowering better decision-making and insight into the possible future contexts. [18].

3. Data mining architecture

A large data mining system architecture can be made up of the following primary components, as illustrated in Figure 2:

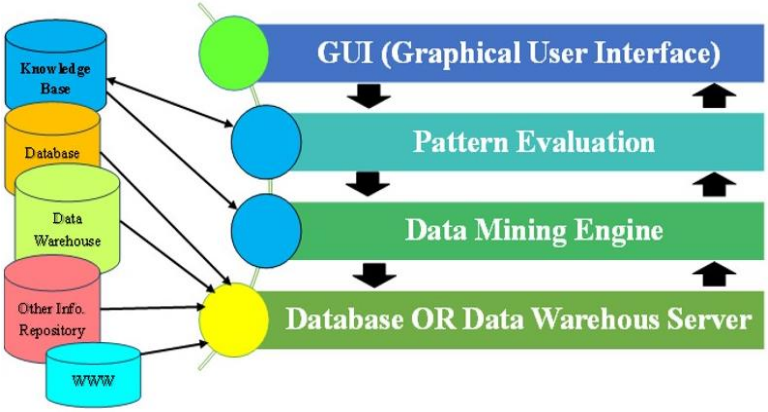

Fig. 2: Architecture of a Standard Data Mining System.

Here's a breakdown of the typical Data Mining Architecture:

a) Data Sources

- Raw Data: Data comes from a variety of sources, such as databases, data warehouses, flat files, web servers, and external sources (e.g., social media, IoT devices).
- Types of Data: Structured, unstructured, and semi-structured data are handled at this stage.
- Data Preprocessing: Raw data is cleaned, integrated, and transformed. Missing values are handled, inconsistencies are removed, and data is formatted to ensure it's ready for mining.

b) Data Warehouse/Database

- A Data Warehouse is designed to hold extensive historical data collected from multiple sources in an organized manner, facilitating data mining. The architecture of the database, such as relational databases, guarantees that the information is readily accessible and enables effective querying.
- ETL (Extract, Transform, Load) process is often used to transfer data from source systems into the warehouse.
- OLAP (Online Analytical Processing): Sometimes used for multidimensional analysis and querying.

c) Data Preprocessing

This step is essential to ensure the quality of the data:

- Data Cleaning: Handling missing values, noise, and inconsistencies in the data.
- Data Integration: Combining data from different sources into a single repository.
- Data Transformation: Normalization, aggregation, and feature extraction to ensure that the data is in a suitable format for analysis.
- Data Reduction: Reducing the size of the data by eliminating irrelevant or redundant features (e.g., dimensionality reduction, feature selection).

d) Data Mining Engine

- The Data Mining Engine is the core of the architecture and performs the actual mining. It includes:
- Algorithms: Various algorithms for different types of data mining tasks, such as classification, clustering, regression, association rule mining, etc.
- Model Building: The mining engine builds models based on the data and the type of mining task (e.g., decision trees, neural networks, clustering models).
- Pattern Discovery: Detects patterns, trends, relationships, and anomalies from the processed data.
- Evaluation: The performance of the model is evaluated based on criteria like accuracy, precision, recall, and F1 score (for predictive tasks).

e) Knowledge Base

The Knowledge Base stores the knowledge acquired from previous mining tasks, along with metadata, algorithms, and evaluation models. It is essential for:

- Storing patterns discovered from prior data mining tasks.
- Retrieving previously found patterns and using them for new analyses or predictions.
- Serving as a source of reference for rule-based systems or for understanding patterns in context.

f) Pattern Evaluation

After the data mining process, this step is responsible for evaluating the discovered patterns:

- Accuracy Assessment: Measures how well the mined patterns or models match real-world data.
- Validation: Verifying if the discovered patterns are reliable and meaningful.
- Interestingness: Measures whether the discovered patterns are useful and actionable. This may include evaluating the significance or novelty of the patterns.

g) User Interface

The User Interface provides an interactive way for the end-user to interact with the system:

- Query Interface: Allows users to query the data or request specific analyses.
- Visualization Tools: Graphical representation of the patterns discovered, often using charts, graphs, or dashboards for easy interpretation.
- Report Generation: Generating reports or summaries that provide insight into the discovered patterns, trends, or anomalies.

h) Decision Support System

- The final step is where the mined knowledge and discovered patterns are used to inform business decisions, strategies, or predictions.

Decision Support Tools: These tools are used to assist decision-makers in interpreting and applying the mined data to real-world scenarios [19].

4. Knowledge management

4.1. Definition of knowledge management

Knowledge management has different definitions. In this paper, we have utilized the definition of knowledge management (KM) as defined by McInerney (2002), declaring it an effort which seeks to improve valuable knowledge in an organization. This can be done by means of building up communication, providing learning opportunities, and facilitating sharing of knowledge artifacts of relevance. This view emphasizes the significance of interaction in knowledge management and learning in organizations [20]. Knowledge management process entails knowledge transfer as well as creation, sharing, and transfer of information. Information technology will be critical in aiding all aspects of knowledge, such as knowledge capture, creation, sharing, acquisition, and utilization. The technologies are essential in efficient knowledge management and are the most important means of its implementation. [21], [22]. Therefore, successful knowledge management necessitates technology that enhances communication and collaboration while improving the capture, sharing, distribution, and application of knowledge, as illustrated in Figure 3:

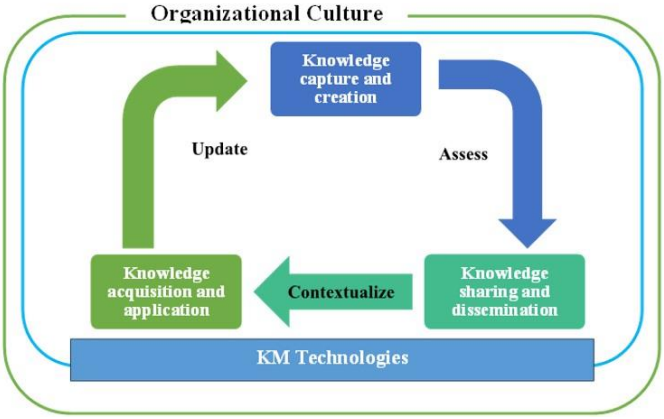

Fig 3: Knowledge Flows and the Process of Creation, Sharing, and Distributing Knowledge.

4.2. Knowledge resources

reveal hidden relationships between KM and organizational performance, facilitating more effective KM implementation.

- Collaboration and Teamwork: Each worker's logs and documents were examined to understand their behavior and build a knowledge flow among workers. Data mining techniques can analyze and develop prototypes of group-based knowledge flows (GKFs) tailored for task-oriented groups.
- Construction Industry: A significant portion of industry information exists in textual data formats. This situation leads to the application of text mining techniques to manage textual information for discovering and managing industrial knowledge [24].

5. Data Mining Aids Health Care

- Each day, the healthcare sector creates significant quantities of complicated and extensive information related to patients, hospital assets, disease diagnoses, electronic medical records, medical equipment, and more.
- A wealth of data serves as essential resources that can be processed and examined to extract knowledge, which helps in saving costs and making decisions. Data mining identifies a range of tools and methods. That can be used on this processed information to uncover hidden patterns and, assist healthcare workers by providing additional insights for decision-making [25]. Professionals of health care are the only persons who make decisions as shown in Figure 4:

Fig. 4: Basic Cycle for the Works Best in the Hospital.

5.1. Health informatics divisions

Health informatics can reasonably be categorized into four primary areas:

- Administration of health services.
- Clinical care.
- Medical research.
- Training [26].

The next sections provide a summary of every division and describe how data mining is currently being used or could be used to enhance and develop each area:

5.1.1. Administration of health services

Healthcare managers are faced with numerous important choices daily. Like in any management role, the effectiveness of these choices is closely tied to the reliability of the information they rely upon. For instance, those running hospitals need to determine the quantity of supplies required, the number of staff members needed, and the amount of available beds for the following month [27]. Managers need to accurately forecast how many patients will come in the following month and how long each patient is likely to remain in the hospital to make this choice. Similarly, health officials at both federal and provincial levels have to determine if an epidemic is happening, and if it is, which preventive strategies will work best to fight it. To reach these conclusions, the administration requires a system that can reliably predict an outbreak and evaluate the costs and advantages of different preventive strategies [28].

5.1.2. Clinical care

Diagnostic choices are solely the responsibility of healthcare professionals such as doctors and nurses, who determine treatment options based on a range of elements including the patient's medical history, imaging studies, lab results, and other forms of patient information, whether textual or multimedia. The medical informatics field improves the capacity of the healthcare providers to retrieve vital information within a short period of time thus making them make more informed decisions. As an illustration, a centralized medical record database enables the doctors of the local clinics to access all the pertinent health information concerning patients irrespective of their physical location in the country. Moreover, the use of data mining methods on this centralized database will provide clinicians with analytical and predictive aspects that they would otherwise not have identified with ease

based on the available data. On the same note, predictive modelling can help the health care practitioners to determine whether a patient would receive better care in an outpatient or inpatient [29].

5.1.3. Medical research

The most successful applications of data mining to health informatics can now be found in the medical research. This tendency is caused by the reality that the health-related data is often stored in small sets of data that are distributed by various clinics, hospitals, and research institutes. However, a significant number of data mining programs that assist in clinical and administrative decision-making require system-atic and centralized data collection. On the other hand, data mining methods can be relatively used to small and fragmented datasets to allow the researcher to discover the relationship between causes and effects, develop valuable models and predictive scoring models using the available data [30].

5.1.4. Education and training

The fourth area of medical informatics appears to be dedicated to educating new health care professionals, as well as the continuous education and improvement of the knowledge of the already existing staff members due to the recent changes in technologies. This is the area of medical informatics which can be reconsidered as a part of the growing trend in online education. Online learning data mining methods have been on the rise of late with early results showing potential results. Increasing attention is paid to the application of data mining strategies in online learning, and the initial experience shows positive outcomes. These methods can benefit all three stakeholders in an educational setting: students, instructors, and administrators. They are capable of monitoring student performance across various tasks and recommending appropriate resources, materials, and pathways to improve the overall educational experience [31]. Educators can use data mining methods to get unbiased insights into how a course is organized and what it includes. These techniques can show how students learn and can help categorize them into smaller groups based on shared learning styles and requirements [32].

Managers use data mining methods to understand how users behave. This helps them improve servers, manage network traffic, and assess how well the educational programs they provide are working [33]. The two case studies that follow offer an overview of a data mining approach and a comparatively recent health informatics software e-learning tool called HOMER. locate articles related to a specific gene [34].

6. Application of Data Mining in Health Care Sector

Companies and marketing organizations have the potential to lead the healthcare sector in utilizing data mining techniques to derive valuable insights from information. There have been successful implementations of data mining applications within the healthcare domain, three examples of which are outlined below:

- Infection Control in Hospitals Each year, around 2 million patients in the United States suffer from infections they acquire while in the hospital, and the cases of infections that resist treatment have reached alarming heights. To spot outbreaks early and identify resistance, it is essential to actively monitor the situation. Studies on computerized monitoring have concentrated on finding patients at higher risk, using expert systems, possible causes, and spotting unusual occurrences of set events. At the University of Alabama, a monitoring system utilizing data mining methods has been set up to discover new and significant patterns in infection control data. This system applies association rules to patient care and culture data sourced from laboratory information management systems and produces monthly reports that an infection control specialist reviews. The developers of this system assert that enhancing infection control through their data mining approach is much more sensitive and considerably more precise than conventional infection control tracking.
- Ranking Hospitals Health organizations assess hospitals and insurance companies based on data that healthcare providers submit. While these submissions are meant to adhere to a uniform format, studies indicate that there is potential for improvement. Techniques for data analysis have been utilized to investigate how reporting is done. By applying the codes from the International Classification of Diseases, 9th Revision, which include risk factors, and creating patient profiles, it is possible to perform cluster and association analyses to reveal how risk factors are expressed. Having a standard way to report is essential since hospitals that disclose risk factors may predict lower patient death rates. Even though their success is similar to that of other hospitals, they could be ranked lower because they pointed out a greater gap between expected and actual death rates. A standardized reporting method would also be crucial for making valid comparisons across different hospitals.
- Recognizing patients at high risk, American Health Ways offers diabetes management services for healthcare facilities and insurance plans aimed at enhancing treatment quality and lowering expenses for individuals living with diabetes. To improve the identification of high-risk patients, American Health Ways employs predictive modeling technology.

Extensive patient information is combined and aggregated to predict the likelihood of short-term health problems and proactively intervene for better short- and long-term outcomes.

Patients at risk for a high-risk disease are identified by a potent data mining and modeling solution. This gives nurse care coordinators a head start. begin by figuring out which patients are at high risk so that measures can be implemented to raise the bar for patient treatment and avoid potential health issues in the future [35].

6.1. Data mining techniques in healthcare

In data mining, analytical techniques are generally utilized, which have historically been labeled as mathematical methods and algorithms. Although data mining is considered a contemporary technology, the practice of data analysis has long been established [36]. The reason these methods relate to large databases is due to the affordability of storage and increased processing capabilities [37].

Numerous data mining techniques, such as artificial neural networks, decision trees, genetic algorithms, and the nearest neighbor method, have demonstrated their efficacy in the healthcare sector [38].

- Artificial Neural Networks

Analytical methods referred to as artificial neural networks are inspired by how the human brain learns. The brain, after going through a learning process, can create hypotheses from previous experiences, which is similar to how neural networks can forecast changes and events in a system following their training. These networks consist of interconnected input and output units that have a distribution of weights referred to as neural networks. The training occurs through a balancing system that focuses on the relationships between the data

samples provided. Depending on the strength of the cause-and-effect relationships among certain data points, connections between "neurons" are made stronger or weaker. As a result, the constructed network is prepared to accept new information and react based on what it has previously learned. Artificial neural networks work well in multiprocessor environments, where many operations are executed simultaneously [39].

- The nearest neighbour method

This method is also applicable for data classification. In contrast to other approaches, it eliminates the need for a training phase to develop a model; instead, the data itself acts as the model. When new data is introduced, the algorithm examines the entire database to identify a collection of examples that most closely match the specified criteria, utilizing this information to forecast the result [40]. The research carried out using the nearest neighbour approach on a standard data collection to evaluate how well heart disease diagnoses are performed showed that this method reached an accuracy level of 97.4%. This percentage surpasses all other previously published studies that used the same data [41].

In healthcare, data mining requires strong collaboration between healthcare quality managers and data mining specialists. It includes data-based assessments and evaluations driven by particular interests.

Analysis motivated by interest can be categorized into seven essential stages:

1) Gaining insights from the relevant field.
2) Developing business-related inquiries.
3) Analyzing these business inquiries.
4) Converting business questions into data mining queries.
5) Implementing data mining queries.
6) Analyzing the outcomes of data mining efforts.
7) Interpreting the results from data mining to derive answers.

Analysis that relies on data is often preferred since it can identify surprising patterns in information. Typically, association rules are applied in this kind of analysis. Combining both analysis types has its advantages and disadvantages. On one hand, users might feel overwhelmed by too many results that exceed what they anticipated. On the other hand, unforeseen trends might not be effectively recognized [42].

6.2. Benefits of health information technology applications

Electronic medical records have been enabled by the use of tracking of patient visits. The information includes multiple elements, including demographics of patients, updates on treatments, observations made in the examination, medications, medical history, laboratory tests, and so on. The information system can help simplify and automate the operations in the healthcare facilities. Nonetheless, patient confidentiality and ethical aspects in data mining are major challenges in medical data mining. In order to make the data mining more accurate, one must create a considerable amount of documentation. In as much as medical records are sensitive information, their utilization may help in treating serious health conditions. Before undertaking the data mining, the healthcare organizations should first have a clear policy on patient information privacy and security. This should be an effective policy that guarantees the confidentiality of patients [43]. Data mining tools can be applied in healthcare organizations in various areas. As an example, physicians can use models that evaluate the clinical and quality indicators, customer satisfaction level, and the economic factors. This complex strategy gives the opportunity to assess the practice of physicians, which makes it easier to allocate resources, make it more cost-effective, and make decisions using available evidence. It also helps in the identification of high-risk patients to be intervened on time and the overall healthcare provision can be optimized [44].

The integration of data mining in the health information systems enables the medical institutions to reduce bias in their decision-making processes and acquire useful information about medicine. Predictive models are an essential tool to the healthcare workers as it helps them improve their knowledge and expertise. Data mining determines the main diseases and conditions that are being observed through a process known as predictive modeling. This involves examining the patient records and prescription of a medical practitioner and identifying the most common problems with patients. Healthcare prediction can be divided into two steps, namely the learning step and the decision step. In the learning phase, the large amount of data is reduced to a smaller and simpler amount of data. The characteristics and items of this new set are much less than those of the original dataset. These instructions are then applied in making informed decisions at this step [45]. A prediction algorithm is used to make predictions using a new data set that is prepared based on new cases whose results are unknown. This algorithm examines the features of a new item by examining its similarities with the items in the chosen data set. After identifying a relation, the new item is assigned the result similar to the average result of the corresponding item. In the medical predictive data mining, we want to have a clear predictive model which will give us accurate predictions which are useful in assisting the doctors to improve their prognosis, diagnoses and treatment plans. The main factors to consider are: is there sufficient data and predictive variables to build a model with an acceptable level of performance; how much association is present between the variables and the outcome; is there significant correlation between the variables or is it possible to establish immediate factors to be obtained through the original variables so as to enhance the performance of the predictive models [46]. The sample usage of data mining is analysis of biomedical signals representing internal rules and reaction to different stimuli. This methodology is highly useful when the information about the interactions between various subsystems is limited, and methods of traditional analysis are not sufficient, particularly when it is related to nonlinear associations. Data mining is a link between the inferences made out of the unremitting data, e.g., biomedical signals measured on patients in intensive care units to the creation of an intelligent monitoring setup that sends alerts, notifications, and alarms in case of severe pre-determined events. The use of association rules entails identifying either all or several key rules within primary subsets that categorize information as either a cause or an effect. This type of inquiry holds considerable interest for healthcare professionals aiming to uncover correlations among diseases, lifestyles, demographics, and survival rates relative to treatments. Matching tasks are employed to substantiate claims regarding the implementation or removal of specific rules within the knowledge model [47].

The responsibilities of managers who are in charge of the quality of healthcare services may be summarized as improving clinical procedures, the medical aspects, and the cost/benefit ratio, as well as administrative quality. The quality of data, standards, plans, and the process of healthcare quality management are all important considerations.

Quality managers can use data mining to complete the following tasks:

1) Finding fresh theories for quality indexes for data, standards, designs, and therapies.
2) Verifying the validity of the provided quality indices for data, standards, plans, and procedures.
3) enhancing, reinforcing, and modifying the indicators of data quality, norms, plans, and therapies.
4) If the current knowledge in the field is given careful consideration during the data mining process, data mining can be used to aid with these activities [48].

6.3. Application challenges

It's both difficult and intriguing to use data mining, knowledge discovery, and machine learning approaches to medical and healthcare data. The quality of the data might vary, but it is often massive, complicated, diverse, and hierarchical. Prior to exploration and discovery, data must be preprocessed and transformed. The data features may occasionally be unsuitable for analysis or exploration. Data conversion is the problem at hand. Prior to any lesson or exploration, convert it to a suitable format [49]. Before data mining may start, a number of problems must be fixed. In the following, we discuss some of the difficulties encountered during data mining in medical databases. Large amounts of data because there are large amounts of medical databases, existing data mining tools might need to take a sample from the database. Another way is to choose specific features from the database [50]. In both methodologies, expertise in the subject matter can be leveraged to discard unnecessary data or characteristics, thereby minimizing the database size. Medical databases are consistently updated with new laboratory test results and patients' ECG records. Consequently, any data mining technique must be able to continuously refresh the insights gained. Inconsistent Data Representations Errors that occur during data entry frequently result in inconsistencies, which are common challenges faced in this context [51]. Inconsistencies can arise in data representation when multiple patterns are available to convey a particular meaning. An example here would be in describing the position of colitis, one system may be 20 cm or 30 cm in the other a description of the length of the sigmoid or rectum. Such difference in data representation may result in wrong diagnosis or treatment because similar concept on the site of colitis is being modeled differently in different systems. Also, the nature of data presented does not always correspond to the data type presented. An example is a column of data that is categorical (numerical) but it is a nominal or ordinal variable that is represented by numbers instead of a continuous variable. This discrepancy is significant for statistical analysis, “mean and variance”.

In the data preparation, modeling, and model evaluation stages, one must take into account prior knowledge (background), such as concept hierarchy, domain expertise, prior knowledge, and rule patterns. There are several forms in which background knowledge may be represented, including examples from the fields of decision rules, Bayesian models, fuzzy sets, and conceptual hierarchies. Background knowledge must be taken into account throughout the KDD process, as illustrated in Figure 5.

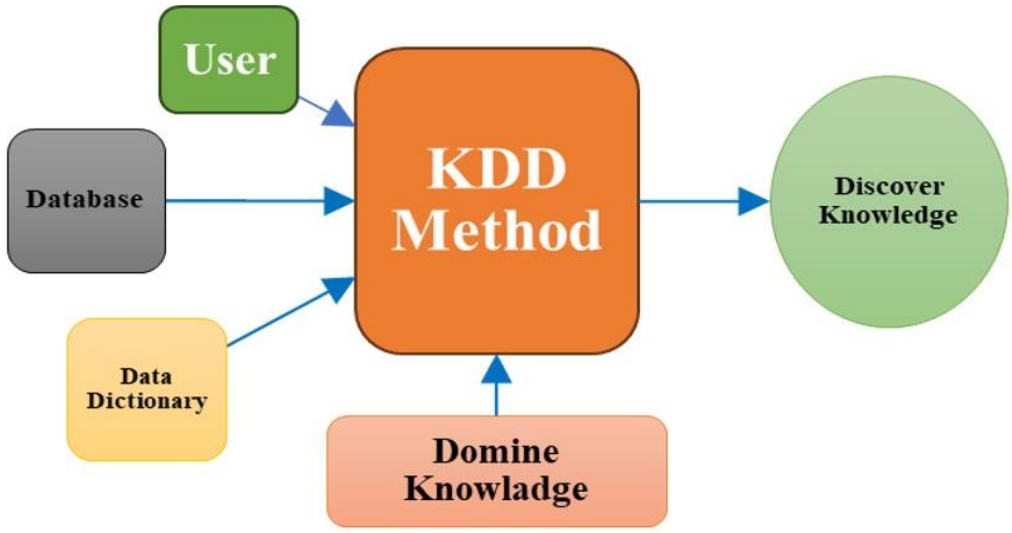

Fig. 5: Background Knowledge Role in KDD Process.

Inadequate Integration Health information is scattered and divided among hospitals, insurance firms, and governmental agencies. This creates a significant obstacle for merging data and analyzing it, especially in terms of the reliability of the outcomes and the meaning of a generated rule. To combine data from different systems, a shared data dictionary and standardized procedures can be used. The rise of XML as a data standard is becoming more popular, simplifying the integration process [53].

The computational complexity of certain data mining techniques does not rise linearly with the increasing number of variables. As the number of variables grows, the time required can become impractical. Techniques such as principal component analysis, which is available in the XLMiner data mining tool, assist in reducing the dataset's variable count while preserving much of the original variation. Additionally, having a clear understanding of the specific domain can facilitate the elimination of superfluous attributes from the data mining workflow [54], [55].

Lacking or Partial Information Clinical database systems frequently fail to gather all necessary data for analysis or exploration. Certain data components are overlooked because of omission, lack of relevance, increased risk, or because they do not apply to a certain clinical situation. Learning techniques like logistic regression may need a full set of data components. Although some methods allow for missing information, the data that was not gathered might still offer valuable insights and should be considered. One way to manage the absent data is to replace missing values with those that are deemed most likely [56], [57].

Medical databases often contain various forms of noise. Consequently, it is important for data mining techniques to exhibit reduced sensitivity to such noise [58].

Amount of Findings The number of results generated by various data mining techniques is overwhelming. In the areas of hospital infection management and public monitoring, as well as in data related to Septic Shock patients, association rule mining has been applied. Both these projects revealed an excessive number of rules. Additionally, there are challenges such as the presence of simple and closely related patterns in data about drug reactions and chronic hepatitis [59], [60].

7. Current Gaps

Here is a clear, concise summary of the current gaps and challenges at the intersection of advanced software systems and data mining in healthcare:

7.1. Privacy, security, and ethical constraints

- High risk of privacy breaches, re-identification, and misuse of sensitive health data.
- Conflicting requirements between innovation and strict regulations (HIPAA, GDPR).
- Ethical concerns around algorithmic bias, fairness, and lack of transparency.
- Limited patient control over data use; governance frameworks remain immature.[61]

7.2. Technical and algorithmic challenges

- High-dimensional, multimodal, and non-IID data make modeling difficult.
- Deep learning models often lack the interpretability required for clinical trust.
- Handling temporal, dynamic patient trajectories remain difficult.
- Scaling mining algorithms to real-world, high-volume healthcare data is nontrivial.[62]

7.3. Organizational and human factors

- Many institutions lack technical expertise, data engineers, and AI-literate clinicians.
- Resistance to adoption due to mistrust of "black-box" tools or workflow disruption.
- Limited budgets and outdated infrastructure slow adoption of advanced mining systems.[63]

7.4. Governance, accountability, and transparency issues

- Lack of standardized frameworks for model auditing, monitoring, bias detection, and lifecycle management.
- Poor documentation (e.g., dataset provenance, model cards, validation history).
- Ambiguity around accountability when AI-driven decisions contribute to errors.
- Need for audit trails, logging, and post-deployment evaluation remains unmet.[64]

8. Software Architecture and Typical Challenges for Healthcare Data Mining

Below is a structured architecture broken into seven layers, each with its role, technologies, and typical challenges (Figure 6):

8.1. Layer 1: Data Sources (Hospital Ecosystem)

A. Clinical & administrative systems

- EHR/EMR (Epic, Cerner, Allscripts, Meditech)
- LIS (Lab Information Systems)
- RIS/PACS (Radiology systems; DICOM images)
- Pharmacy systems

B. Medical devices & IoMT

- Bedside monitors
- Wearables (heart rate, glucose sensors)
- Smart infusion pumps
- Ventilators, telemetry systems

C. External & public data

- Genomics repositories
- Claims databases
- Public health datasets
- Clinical trials systems [65].

8.2. Layer 2: Ingestion & Interoperability Layer

This is the "pipework" connecting hospital systems to your analytics platform.

A. Ingestion connectors

- HL7v2 listeners
- FHIR REST API ingestion
- DICOMweb for imaging
- IoMT device gateways (MQTT, CoAP, BLE)

B. Technologies

- Apache NiFi (fine-grained flow control + provenance)
- Kafka / Pulsar (streaming ingestion)

- HAPI FHIR servers
- Dicom Router / Orthanc / dcm4che

C. Responsibilities

- Normalize message formats
- Validate schemas
- Apply de-identification when required
- Guarantee delivery (retry, idempotency)

Challenge: HL7 messages and EHR APIs vary greatly between hospitals → require custom adapters [66].

8.3. Layer 3: data lake & warehouse (storage + governance)

A dual-layer storage architecture:

A. Raw data lake ("bronze")

- Stores raw HL7 messages, FHIR bundles, logs, imaging, IoMT streams
- On-prem: Ceph, MinIO, HDFS
- Cloud: S3, GCS, Azure Blob

B. Curated clinical data warehouse ("silver + gold")

- Structured, queryable data mapped to clinical ontologies
- Modeling paradigms:
- OMOP CDM
- FHIR-native warehouse
- Dimensional models (star schema)

C. Data governance layer

- Data catalogs (Amundsen, DataHub)
- Lineage (Apache Atlas, OpenLineage)
- Sensitive data tagging (PHI flags, masking policies)
- Audit trails (required for HIPAA/GDPR)

Challenge: Maintaining consistent terminology mapping — SNOMED ↔ ICD-10 ↔ LOINC [67].

8.4. Layer 4: processing & transformation (ETL/ELT pipelines)

This layer turns raw clinical data into analytics-ready datasets.

A. Batch processing

- Apache Spark
- Databricks
- Flink (batch mode)
- dbt transformations in SQL warehouses

B. Streaming processing

- Kafka Streams, Flink, Spark Streaming
- Real-time anomaly detection (e.g., vitals, ICU alerts)

C. Semantic normalization

- Ontology mapping (SNOMED, RxNorm, LOINC)
- FHIR → relational transformation
- DICOM image metadata extraction

D. Feature engineering layer

- Feature Store (Feast, Tecton) shared across teams
- Provides versioned, validated ML-ready feature sets

Challenge: Guarantee reproducibility of derived data for clinical audit or FDA review [68].

8.5. Layer 5: ML/AI, data mining, and advanced algorithms

Now we reach the actual "data mining" engines.

A. Model training environment

- Kubeflow, MLflow, SageMaker, Vertex AI
- GPU/TPU clusters
- Experiment tracking + reproducibility

B. Common analytical workloads

- Predictive analytics (mortality risk, readmissions, sepsis onset)
- Clustering / patient stratification
- Association mining (comorbidity patterns)
- Process mining (clinical pathways, bottlenecks)
- Temporal data mining (longitudinal lab trajectories)

C. Specialized packages

- NLP: clinical notes processing (cTAKES, MedSpaCy, transformer models)
- Imaging: MONAI, 3D CNNs
- Genomics: GATK, deep variant models
- Time series: TSFresh, deep sequence models

Challenge: Most hospital data are non-IID, messy, and temporally irregular → requires domain-specific modeling [69].

8.6. Layer 6: model serving, microservices, and real-time APIs

Operationalizing models for clinicians.

A. Microservice architecture

Each model gets its own:
- inference API (REST/GRPC)
- monitoring dashboard
- versioning policies
- explainability module (SHAP, LIME)

B. Infrastructure

- Kubernetes clusters
- Istio service mesh (mTLS, traffic routing)
- Knative for serverless inference
- Redis or Feast for real-time features

C. Deployment Patterns

- Batch predictions → nightly risk scores
- Real-time API → EHR calls model for active patients
- Edge inference → wearable or bedside monitor

D. Observability

- Drift detection
- PHI-aware logs
- Error tracking
- Model auditing

Challenge: Low latency + high privacy + high reliability = technically demanding [70].

8.7. Layer 7: applications & end-user integration

A. Clinical decision support (CDS)

- Embedded in EHR (Epic BestPractice Advisories, SMART-on-FHIR apps)
- Alerts, recommendations, risk scores

B. Dashboards & BI tools

- PowerBI, Tableau, Looker
- Hospital operations dashboards
- Real-time bed occupancy, throughput, patient flow analytics

C. Research portals

- De-identified cohort selection platforms
- Self-service analytics for clinical researchers

D. Patient-facing applications

- Personalized care recommendations
- Remote monitoring apps
- Digital therapeutics

Challenge: High risk of alert fatigue that mean requires tight integration with clinical workflow [71].

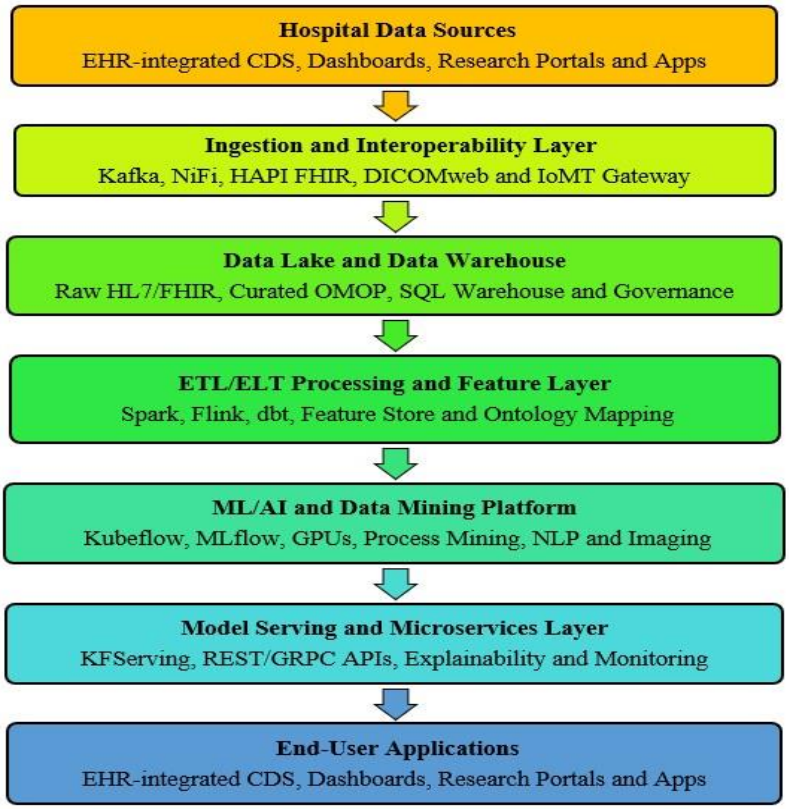

Fig. 6: Complete Architecture Diagram.

9. Engineering Pain Points

1) Data heterogeneity: inconsistent formats, codes, and modalities.
2) Interoperability: bridging legacy systems with modern APIs.
3) Data governance: versioning, lineage, consent tracking.
4) Scalability: high-volume imaging + real-time IoMT streams.
5) Security compliance, such as HIPAA/GDPR, adds strict constraints.
6) MLOps complexity: clinical-grade reproducibility and safety.
7) Hybrid deployments: cloud + on-prem + edge coordination.
8) Reliability: healthcare systems must be 24/7 fault-tolerant [61], [72], [73].

10. Modern Healthcare Data Mining Techniques

The last five years have seen healthcare AI improve at an unparalleled rate because of the emergence of deep learning architectures, the development of large language models (LLMs), privacy-preserving collaboration methods, and an increased desire to explain how AI was used. A narrow detail and a recent summary are given below.

10.1. Machine learning–based techniques

a) Supervised Learning

Used when labeled data exists (e.g., diagnosis codes, outcomes).

Classification algorithms:

- Logistic Regression.
- Random Forests.
- Gradient Boosting (XGBoost, LightGBM).
- Support Vector Machines.

- Neural Networks.

b) Self-supervised Learning (SSL)

A major shift occurred from supervised to self-supervised training, addressing the chronic shortage of labeled medical data.
Popular SSL paradigms include:

- Contrastive learning (e.g., SimCLR variants).
- Masked autoencoders for imaging.
- Predictive coding for physiological signals.
- Impact: High performance with 10–100× fewer labels, enabling broader deployment in smaller hospitals.

Multimodal Deep Learning

c) Systems now combine:

- Imaging.
- Clinical notes.
- Lab values.
- Genomics.

Models like multimodal transformers (image + text) support richer clinical decision support.

10.2. Deep learning

a) CNNs for medical imaging (X-ray, MRI).
b) RNNs/LSTMs & Transformers for time-series data and clinical text analysis.
c) LLMs for summarizing EHRs and extracting clinical information.

10.3. Natural language processing (NLP)

Processes unstructured clinical notes for coding, diagnosis extraction, and decision support.

10.4. Time-series mining

Analyzes continuous monitoring data (ICU vitals, wearables) to detect deterioration and forecast events.

10.5. Genomic and bioinformatics techniques

Deep learning, GWAS, and graph algorithms enable personalized medicine and mutation detection.

10.6. Graph-based mining

Graph neural networks and medical knowledge graphs model relationships between diseases, drugs, and patients.

10.7. Privacy-preserving mining

Federated learning, differential privacy, and encrypted computation enable multi-center analytics without sharing raw data.

10.8. Anomaly detection

Used for fraud detection, rare disease identification, and monitoring abnormalities in clinical signals.

10.9. Reinforcement learning

Optimizes treatment plans, ICU support strategies, and hospital resource management [74].
So, Table 1 displays Common Healthcare Use Cases.

Table 1: Common Healthcare Use Cases

Use Case	Data Mining Techniques	Outcome
Clinical decision support	ML, deep learning, NLP	Diagnosis & treatment recommendations
Population health	Clustering, regression	Risk stratification
Hospital operations	Predictive modeling	Staff scheduling, resource optimization
Imaging diagnostics	CNNs	Automated detection (tumors, fractures)
Drug discovery	GNNs, association mining	Drug interaction prediction

11. Conclusion

Data mining has found successful applications in well-regarded fields like business, marketing, and retail. As a result, it is now relevant in knowledge discovery in databases (KDD) across various industries and economies. Its significance and utilization are particularly increasing in the fields of medicine and public health. The medical field has immense prospects under data mining since the healthcare systems can use data and analytics more systematically. By doing so, one can easily identify areas of inefficiency and the best practices that can be employed to improve the quality of care besides reducing costs. Data mining tools are useful in the medical sector to collect useful information. Even though it is very challenging to predict the diseases using data mining applications, it saves the number of human labor required and enhances the accuracy of the diagnosis. It might result in cost and time savings when it comes to the manpower and expertise required to create efficient data mining tools to be applied in particular applications. Nevertheless, medical data analysis may be a dangerous

task because it may contain noisy, irrelevant, and voluminous data. The success of data mining procedures depends a lot on the kind of dataset that is used in testing.

The possibilities of data mining in healthcare with the help of state-of-the-art software concepts and knowledge extraction methods have a major potential to transform the contemporary medical practice. The combination of structured and unstructured data including EHR records and medical images and ge-nomic sequences allows creating predictive methods and intelligent systems that can enhance the work of diagnostics, clinical workflow optimization, individualized treatment, and cost-reduced operations.

To conclude the analysis, the four key pillars that are required to assure successful implementation are:

a) Well-developed Data Infrastructure - that can combine various data sources and ensure that it integrates safely with high support of the interoperability criteria and data privacy laws.
b) Sophisticated Analytics and Machine Learning Models - these are built to identify clinically relevant trends and aid in decision making.
c) Knowledge Representation and Explainability Frameworks - will the insights mined be interpretable, actionable and aligned with clinical reasoning.
d) Ethical, Secure, and Scalable Software Systems - ensuring patient rights and deploying the technology on a large scale throughout health care settings.

Despite the existing difficulties, particularly in the field of data quality, mitigation of bias, interpretability of the models, and their integration into clinical work-flows, the pattern of technological advancement indicates further progress. With increasing use of AI, distributed computing, and real-time analytics in healthcare, data mining will play a more significant role in the realization of precision medicine, population healthcare, and proactive, not reactive care.

In the beginning we need to know what data mining is and how it can be achieved and this is what we introduce in this study. As the second step, we would like to discuss the above research questions, pointing out several overall directions:

- Enhancing the methods of data quality and interoperability.
- Creating explainable, trustworthy, and understanding AI.
- Maybe because of that, the key aspect should be secure analytics pipelines that are ethical.
- Utilizing multimodal and real-time streams of data.
- Developing clinical NLP and reasoning systems.
- Further decision support through the creation of knowledge graphs.
- Learning about human-AI collaboration in clinical practice.
- Population-level health effects and equity: Providing a solution to this challenge.

These add up to a map that a researcher and a practitioner wishing to build the next generation of healthcare data mining systems can use. Finally, the conclusion emphasizes the fact that software engineering principles and knowledge extraction methodology are not only a technical addition but an engine of the future of healthcare. The tools enable clinicians and researchers to transform large datasets into valuable information that can be used to translate computational intelligence into higher patient results and more stable and efficient healthcare infrastructure.

References

[1] S. Khan OKESOLA and M. Shaheen, "From data mining to wisdom mining," Journal of Information Science, vol. 49, no. 4, pp. 952-975, 2023. https://doi.org/10.1177/01655515211030872.
[2] R. Koppel, "Estimating the United States' cost of healthcare information technology," in Healthcare Information Management Systems: Cases, Strategies, and Solutions: Springer, 2022, pp. 3-38. https://doi.org/10.1007/978-3-031-07912-2_1.
[3] E. P. Ambinder, "Electronic Health Records," Journal of Oncology Practice, vol. 1, no. 2, pp. 57-63, 2005, https://doi.org/10.1200/jop.2005.1.2.57.
[4] A. Nambiar and D. Mundra, "An overview of data warehouse and data lake in modern enterprise data management," Big data and cognitive computing, vol. 6, no. 4, p. 132, 2022. https://doi.org/10.3390/bdcc6040132.
[5] D. Shukla, S. B. Patel, and A. K. Sen, "A literature review in health informatics using data mining techniques," International Journal of Software and Hardware Research in Engineering, vol. 2, no. 2, pp. 123-129, 2014.
[6] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," Expert Systems with Applications, vol. 166, p. 114060, 2021/03/15/ 2021, https://doi.org/10.1016/j.eswa.2020.114060.
[7] W. I. D. Mining, Introduction to data mining. Springer, 2006.
[8] A. A. Hasan and H. Fang, "Data Mining in Education: Discussing Knowledge Discovery in Database (KDD) with Cluster Associative Study," presented at the 2021 2nd International Conference on Artificial Intelligence and Information Systems, Chongqing, China, 2021. [Online]. Available: https://doi.org/10.1145/3469213.3471319.
[9] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions," Artificial Intelligence Review, vol. 55, no. 1, pp. 453-563, 2022. https://doi.org/10.1007/s10462-021-10037-9.
[10] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Global Transitions Proceedings, vol. 3, no. 1, pp. 91-99, 2022. https://doi.org/10.1016/j.gltp.2022.04.020.
[11] Y. Yao, "Symbols-Meaning-Value (SMV) space as a basis for a conceptual model of data science," International Journal of Approximate Reasoning, vol. 144, pp. 113-128, 2022. https://doi.org/10.1016/j.ijar.2022.02.001.
[12] S. Chakraborty, S. H. Islam, and D. Samanta, Data classification and incremental clustering in data mining and machine learning. Springer, 2022. https://doi.org/10.1007/978-3-030-93088-2.
[13] O. O., "DATA MINING METHODOLOGY AND ITS APPLICATION," 2021.
[14] D. K. Nayak, A. K. Mishra, and M. Mistry, "A BRIEF STUDY ON DATA MINING AND BIG DATA."
[15] R. R. Asaad and R. M. Abdulhakim, "The Concept of Data Mining and Knowledge Extraction Techniques," Qubahan Academic Journal, vol. 1, no. 2, pp. 17-20, 2021. https://doi.org/10.48161/qaj.v1n2a43.
[16] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," Social Science Research, vol. 110, p. 102817, 2023. https://doi.org/10.1016/j.ssresearch.2022.102817.
[17] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," SN computer science, vol. 2, no. 3, p. 160, 2021. https://doi.org/10.1007/s42979-021-00592-x.
[18] A. Y. Abd Alazeez, "Data mining between classical and modern applications: A review," AL-Rafidain Journal of Computer Sciences and Mathematics, vol. 15, no. 2, pp. 171-191, 2021. https://doi.org/10.33899/csmj.2021.170020.
[19] A. Saini, "Data Mining Architecture–Data Mining Types and Techniques," 2021.
[20] C. McInerney, "Knowledge management and the dynamic nature of knowledge," Journal of the American society for Information Science and Technology, vol. 53, no. 12, pp. 1009-1018, 2002. https://doi.org/10.1002/asi.10109.

[21] D. Harman, "Knowledge Management (KM) Processes in Organizations: Theoretical Foundations and Practice/Harman D," ed: New York, NY: Basic Books, 2011.
[22] A. S. Wahjudewanti, J. H. Tjakraatmaja, and Y. Anggoro, "Knowledge Management Strategies to Improve Learning and Growth in Creative Industries: A Framework Model," Budapest International Research and Critics Institute-Journal (BIRCI-Journal) Vol, vol. 4, no. 2, pp. 1903-1915, 2021. https://doi.org/10.33258/birci.v4i2.1876.
[23] Y. Li, M. Kramer, A. J. Beulens, and J. G. van der Vorst, "A framework for early warning and proactive control systems in food supply chain networks," Computers in Industry, vol. 61, no. 9, pp. 852-862, 2010. https://doi.org/10.1016/j.compind.2010.07.010.
[24] T. Silwattananusarn and K. Tuamsuk, "Data mining and its applications for knowledge management: a literature review from 2007 to 2012," arXiv preprint arXiv:1210.2872, 2012. https://doi.org/10.5121/ijdkp.2012.2502.
[25] S. V. G. Subrahmanya et al., "The role of data science in healthcare advancements: applications, benefits, and future prospects," Irish Journal of Medical Science (1971-), vol. 191, no. 4, pp. 1473-1483, 2022. https://doi.org/10.1007/s11845-021-02730-z.
[26] A. H. Al-Obaidi, "Data Mining In Healthcare Sectors," Wisdom Journal for Studies & Research, vol. 4, no. 06, pp. 1215-1237, 2024. https://doi.org/10.55165/wjfsar.v4i06.502.
[27] A. H. T. Al-Ghrairi, A. A. Mohammed, and H. M. Saeed, "An Application of Web-based E-Healthcare Management System Using ASP. Net," Webology, vol. 18, no. 1, 2021. https://doi.org/10.14704/WEB/V18I1/WEB18089.
[28] A. M. Mosadeghrad, "Factors influencing healthcare service quality," International journal of health policy and management, vol. 3, no. 2, p. 77, 2014. https://doi.org/10.15171/ijhpm.2014.65.
[29] S. A. Alowais et al., "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," BMC medical education, vol. 23, no. 1, p. 689, 2023. https://doi.org/10.1186/s12909-023-04698-z.
[30] M. Javaid, A. Haleem, and R. P. Singh, "Health informatics to enhance the healthcare industry's culture: An extensive analysis of its features, contributions, applications and limitations," Informatics and Health, 2024. https://doi.org/10.1016/j.infoh.2024.05.001.
[31] H. Chang, "Recent movement on education and training in health informatics," Healthcare Informatics Research, vol. 20, no. 2, p. 79, 2014. https://doi.org/10.4258/hir.2014.20.2.79.
[32] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," Ieee Access, vol. 5, pp. 15991-16005, 2017. https://doi.org/10.1109/ACCESS.2017.2654247.
[33] V. Švábenský, J. Vykopal, P. Čeleda, and L. Kraus, "Applications of educational data mining and learning analytics on data from cybersecurity training," Education and Information Technologies, vol. 27, no. 9, pp. 12179-12212, 2022. https://doi.org/10.1007/s10639-022-11093-6.
[34] P. Ryan et al., "Health Outcomes and Medical Effectiveness Research (HOMER): A Systematic Approach to Exploring Hill's Causal Viewpoints in Observational Data," in PHARMACOEPIDEMIOLOGY AND DRUG SAFETY, 2014, vol. 23: WILEY-BLACKWELL 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, pp. 145-146.
[35] E. Shirzad, G. Ataei, and H. Saadatfar, "Applications of data mining in healthcare area: A survey," Engineering & Applied Science Research, vol. 48, no. 3, 2021.
[36] A. Alam and A. Mohanty, "Predicting students' performance employing educational data mining techniques, machine learning, and learning analytics," in International Conference on Communication, Networks and Computing, 2022: Springer, pp. 166-177. https://doi.org/10.1007/978-3-031-43140-1_15.
[37] S. Usman, R. Mehmood, I. Katib, and A. Albeshri, "Data locality in high performance computing, big data, and converged systems: An analysis of the cutting edge and a future system architecture," Electronics, vol. 12, no. 1, p. 53, 2022. https://doi.org/10.3390/electronics12010053.
[38] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," Decision Analytics Journal, vol. 3, p. 100071, 2022. https://doi.org/10.1016/j.dajour.2022.100071.
[39] Z. Zhang and Z. Zhang, "Artificial neural network," Multivariate time series analysis in climate and environmental research, pp. 1-35, 2018. https://doi.org/10.1007/978-3-319-67340-0_1.
[40] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in 2019 international conference on intelligent computing and control systems (ICCS), 2019: IEEE, pp. 1255-1260. https://doi.org/10.1109/ICCS45141.2019.9065747.
[41] M. Shouman, T. Turner, and R. Stocker, "Disease Patients," International Journal of Information and Education Technology, vol. 2, no. 3, 2012.
[42] B. Milovic and M. Milovic, "Prediction and decision making in health care using data mining," Arabian Journal of Business and Management Review (Kuwait Chapter), vol. 1, no. 12, pp. 126-136, 2012. https://doi.org/10.11591/ijphs.v1i2.1380.
[43] A. H. Abdulwahhab, A. H. Abdulaal, A. H. T. Al-Ghrairi, A. A. Mohammed, and M. Valizadeh, "Detection of epileptic seizure using EEG signals analysis based on deep learning techniques," Chaos, Solitons & Fractals, vol. 181, p. 114700, 2024. https://doi.org/10.1016/j.chaos.2024.114700.
[44] A. Sheikh et al., "Health information technology and digital innovation for national learning health and care systems," The Lancet Digital Health, vol. 3, no. 6, pp. e383-e396, 2021. https://doi.org/10.1016/S2589-7500(21)00005-4
[45] C. Sirocchi, A. Bogliolo, and S. Montagna, "Medical-informed machine learning: integrating prior knowledge into medical decision systems," BMC Medical Informatics and Decision Making, vol. 24, no. Suppl 4, p. 186, 2024. https://doi.org/10.1186/s12911-024-02582-4
[46] D. A. Neu, J. Lahann, and P. Fettke, "A systematic literature review on state-of-the-art deep learning methods for process prediction," Artificial Intelligence Review, vol. 55, no. 2, pp. 801-827, 2022. https://doi.org/10.1007/s10462-021-09960-8.
[47] H. Ameri, S. Alizadeh, and E. A. Z. Noughabi, "Application of data mining techniques in clinical decision making: A literature review and classification," Handbook of Research on Data Science for Effective Healthcare Practice and Administration, pp. 257-295, 2017. https://doi.org/10.4018/978-1-5225-2515-8.ch012.
[48] D. McGilvray, Executing data quality projects: Ten steps to quality data and trusted information (TM). Academic Press, 2021.
[49] T. Shaik et al., "Remote patient monitoring using artificial intelligence: Current state, applications, and challenges," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 13, no. 2, p. e1485, 2023. https://doi.org/10.1002/widm.1485.
[50] T. Sarwar et al., "The secondary use of electronic health records for data mining: Data characteristics and challenges," ACM Computing Surveys (CSUR), vol. 55, no. 2, pp. 1-40, 2022. https://doi.org/10.1145/3490234.
[51] A. Ed-daoudy and K. Maalmi, "Breast cancer classification with reduced feature set using association rules and support vector machine," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 9, no. 1, p. 34, 2020. https://doi.org/10.1007/s13721-020-00237-8.
[52] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," Applied artificial intelligence, vol. 17, no. 5-6, pp. 375-381, 2003. https://doi.org/10.1080/713827180.
[53] Z. Wu and V. Trigo, "Impact of information system integration on the healthcare management and medical services," International Journal of Healthcare Management, vol. 14, no. 4, pp. 1348-1356, 2021. https://doi.org/10.1080/20479700.2020.1760015.
[54] W.-T. Wu et al., "Data mining in clinical big data: the frequently used databases, steps, and methodological models," Military Medical Research, vol. 8, pp. 1-12, 2021. https://doi.org/10.1186/s40779-021-00338-z.
[55] G. Shmueli, P. C. Bruce, K. R. Deokar, and N. R. Patel, Machine learning for business analytics: Concepts, techniques, and applications with analytic solver data mining. John Wiley & Sons, 2023.
[56] B. Al-Sahab, A. Leviton, T. Loddenkemper, N. Paneth, and B. Zhang, "Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview," Journal of Healthcare Informatics Research, vol. 8, no. 1, pp. 121-139, 2024. https://doi.org/10.1007/s41666-023-00153-2.
[57] T. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," Malaysian Journal of Computer Science, pp. 132-148, 2022. https://doi.org/10.22452/mjcs.sp2022no1.10.
[58] A. Ampavathi, "Research challenges and future directions towards medical data processing," Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, vol. 10, no. 6, pp. 633-652, 2022. https://doi.org/10.1080/21681163.2021.2018665.

[59] A. Gupta, R. Singh, V. K. Nassa, R. Bansal, P. Sharma, and K. Koti, "Investigating application and challenges of big data analytics with clustering," in 2021 international conference on advancements in electrical, electronics, communication, computing and automation (ICAECA), 2021: IEEE, pp. 1-6. https://doi.org/10.1109/ICAECA52838.2021.9675483.

[60] M. M. Rahman, M. S. Chowdhury, M. Shorfuzzaman, L. Karim, M. Shafiullah, and F. Azzedin, "Enhancing Septic Shock Detection through Interpretable Machine Learning," CMES-Computer Modeling in Engineering & Sciences, vol. 141, no. 3, 2024. https://doi.org/10.32604/cmes.2024.055065.

[61] P. Kaushik, K. Sharma, M. K. Mahawar, J. Wasim, G. Dey, and S. A. Nibiya, "Ethical Considerations in Data Mining and Analytics," in *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, 2024: IEEE, pp. 1516-1521. https://doi.org/10.1109/ICAC2N63387.2024.10895012.

[62] T. Tsalko, S. Nevmerzhytska, S. Krasniuk, S. Goncharenko, and N. Liubymova, "Features, problems and prospects of data mining and data science application in educational management," *Bulletin of Science and Education (Series" Philology", Series" Pedagogy", Series" Sociology", Series" Culture and Art", Series" History and Archeology"),* 2024. https://doi.org/10.52058/2786-6165-2024-5(23)-637-657.

[63] E. I. Kabanov, M. V. Tumanov, V. S. Smetanin, and K. V. Romanov, "An innovative approach to injury prevention in mining companies through human factor management," *Записки Горного института,* no. 263 (eng), pp. 774-784, 2023.

[64] P. Solana-González, A. A. Vanti, M. M. García Lorenzo, and R. E. Bello Pérez, "Data mining to assess organizational transparency across technology processes: An approach from it governance and knowledge management," *Sustainability,* vol. 13, no. 18, p. 10130, 2021. https://doi.org/10.3390/su131810130

[65] Y. Zhong, L. Chen, C. Dan, and A. Rezaeipanah, "A systematic survey of data mining and big data analysis in internet of things," *The Journal of Supercomputing,* vol. 78, no. 17, pp. 18405-18453, 2022. https://doi.org/10.1007/s11227-022-04594-1.

[66] V. S. Naresh and M. Thamarai, "Privacy-preserving data mining and machine learning in healthcare: Applications, challenges, and solutions," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 13, no. 2, p. e1490, 2023. https://doi.org/10.1002/widm.1490.

[67] A. Nambiar and D. Mundra, "An overview of data warehouse and data lake in modern enterprise data management," *Big data and cognitive computing,* vol. 6, no. 4, p. 132, 2022. https://doi.org/10.3390/bdcc6040132.

[68] A. O. Akindemowo *et al.*, "A Conceptual Framework for Automating Data Pipelines Using ELT Tools in Cloud-Native Environments," 2021. https://doi.org/10.54660/.JFMR.2021.2.1.440-452.

[69] A. Jagadish, *Essential Concepts and Techniques of AI & ML*. Academic Guru Publishing House, 2024.

[70] D. K. Pandiya and N. Charankar, "Integration of microservices and AI for real-time data processing," *International journal of computer engineering and technology (IJCET),* vol. 14, no. 2, pp. 240-254, 2023.

[71] G. S. Reddy, "A review of data warehouses multidimensional model and data mining," *Information Technology in Industry,* vol. 9, no. 3, pp. 310-320, 2021.

[72] A. H. T. Al-Ghrairi, A. A. Mohammed, and H. M. Saeed, "An Application of Web-based E-Healthcare Management System Using ASP. Net," *Webology,* vol. 18, no. 1, 2021. https://doi.org/10.14704/WEB/V18I1/WEB18089.

[73] H. A. Ganatra, "Machine learning in pediatric healthcare: Current trends, challenges, and future directions," *Journal of Clinical Medicine,* vol. 14, no. 3, p. 807, 2025. https://doi.org/10.3390/jcm14030807.

[74] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Natural language processing for smart healthcare," *IEEE Reviews in Biomedical Engineering,* vol. 17, pp. 4-18, 2022. https://doi.org/10.1109/RBME.2022.3210270.