DATA MINING IN HEALTHCARE: FROM METHODS TO REAL-WORLD IMPACT

Dr. Sonia Saini
N C Medical College, Israna Panipat
Email ID: bk89wisdom@gmail.com

ABSTRACT

Purpose: Data mining refers to the process of extracting useful patterns, trends, and insights from large sets of data. In everyday life, it powers many of the tools and services we use. Here are some common applications of data mining in everyday life: shopping & retail, online services, banking & finance, healthcare, transportation, entertainment, smart homes & IoT, and scientific research. In this given paper discuss review on data mining on the tools of the medical diagnosis.

Design/Methodology/Approach: By using some steps, we can report the result, like data collection, data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and at the end, knowledge presentation. We have different methods to solve the issue about data, like classification, clustering, regression, Association rule learning, Natural Language Processing (NLP), Neural Networks & Deep Learning and Anomaly Detection.

Findings: Data mining is very effective tool in healthcare disease diagnosis & prediction, treatment effectiveness, clinical decision support systems (CDSS), medical imaging, healthcare management, fraud detection in medical claims, genomics and drug discovery.

Research Limitations: Normally, we face some challenges in data mining as data quality and preprocessing, high dimensionality of data, scalability of algorithms, privacy and security concerns, and interpretability of results.

Managerial Implications: Data mining is very effective tool in healthcare disease diagnosis & prediction, treatment effectiveness, clinical decision support systems (CDSS), medical imaging, healthcare management, fraud detection in medical claims, genomics and drug discovery.

Originality/Value: In this given paper discuss review on data mining on the tools of the medical diagnosis.

Keywords: Natural Language Processing, Neural Networks & Deep Learning and Anomaly Detection, Clinical Decision Support Systems, Data Mining, World Health Organization

Introduction

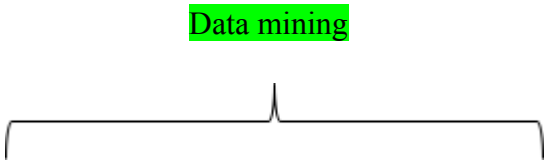
Data mining is the process of finding patterns, identifying recurring trends and relationships within data, detecting anomalies to spotting unusual data points that deviate from the norm, and analysing correlations to determine the relationship between study variables within large datasets to extract meaningful information, actionable insights to transform raw data into information that is used for strategic decision-making. Different type of techniques are used to solve these issues machine learning includes algorithms like classification, regression, and clustering are used to build predictive models, statistical analysis includes methods like regression, hypothesis testing, and descriptive statistics are employed to analyze data and identify trends, database system includes data mining relies on efficient storage and retrieval of large datasets. The impact of the use of data mining is very effective in improving decision-making, increasing revenue, reducing costs, enhancing customer relationships, fraud in retail, healthcare, finance, E-commerce, etc.

Uses of data mining



Business and marketing

In business and marketing, it helps organizations understand customer behaviours, predict trends, increase sales, reduce cost, and improve decision-making, for example, customer relationship management (CRM), Market segmentation, Sales forecasting, product recommendation, pricing strategies, fraud detection, and sentiment analysis.



Business	Marketing
Customer Segmentation	Targeted Marketing
Fraud Detection	Churn Prediction
Market Basket Analysis	Campaign Effectiveness
Risk Management	Trend Analysis
Recommendation System	Sentiment Analysis
	Price Optimization

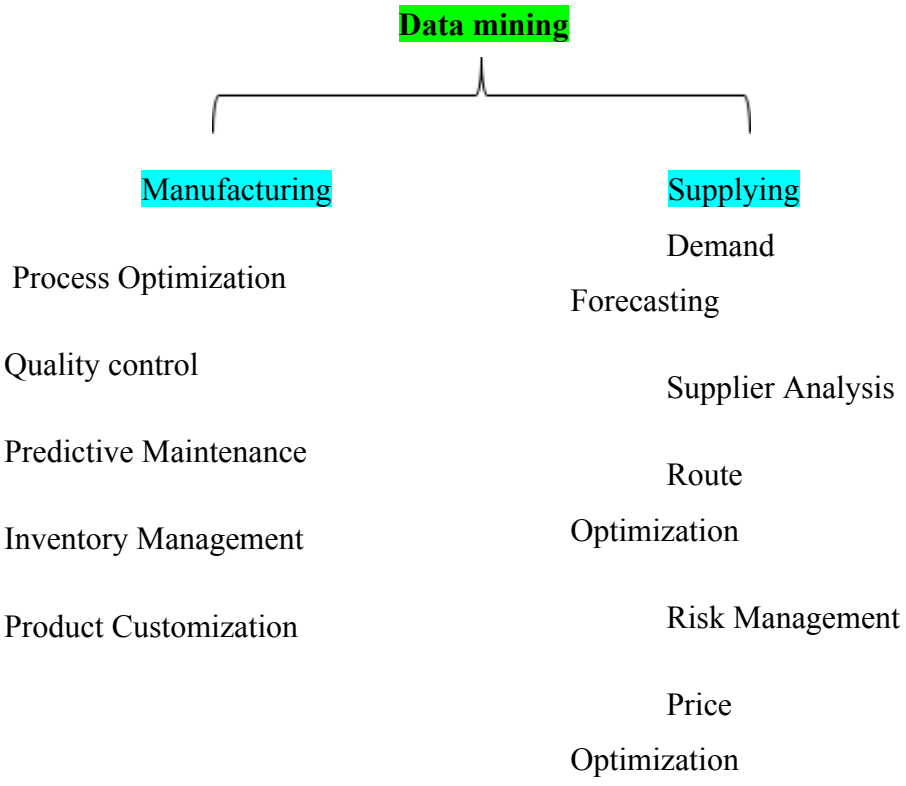
Social Media & Web

Social media and the web involve extracting useful patterns, sentiments, trends, and insights from the vast amounts of data generated by online platforms, such as Facebook, Twitter, Instagram, and LinkedIn. The data can be in different forms like user profiles and demographics, post, comments, hashtags, likes, shares, reactions, follower-following networks, multimedia content (images, videos, stories), clickstream, and engagement data. This is crucial for understanding user behavior, improving services, increasing engagement and conversions, and making data-driven decisions. Web data mining uses automated tools to discover patterns from websites, web logs, clickstreams, and online transactions. The application is very broad in social media, like brand monitoring, trend prediction, influencer identification, and crisis management. On the web, personalized marketing, fraud detection, user experience improvement, and competitor analysis. Different types of tools & technologies are used, like data mining tools, social media analytics tools, web crawling tools, and AI/ML

frameworks. But there are some challenges like data privacy issues (GDPR, data misuse concerns), data volume & velocity (big data handling), Noise in data (irrelevant or misleading content), bias in algorithms (leading to skewed insights). But really, it is very effective and beneficial, such as real-time insight into trends and opinions, improved marketing and engagement strategies, better customer relationship management, and early detection of issues or crises.

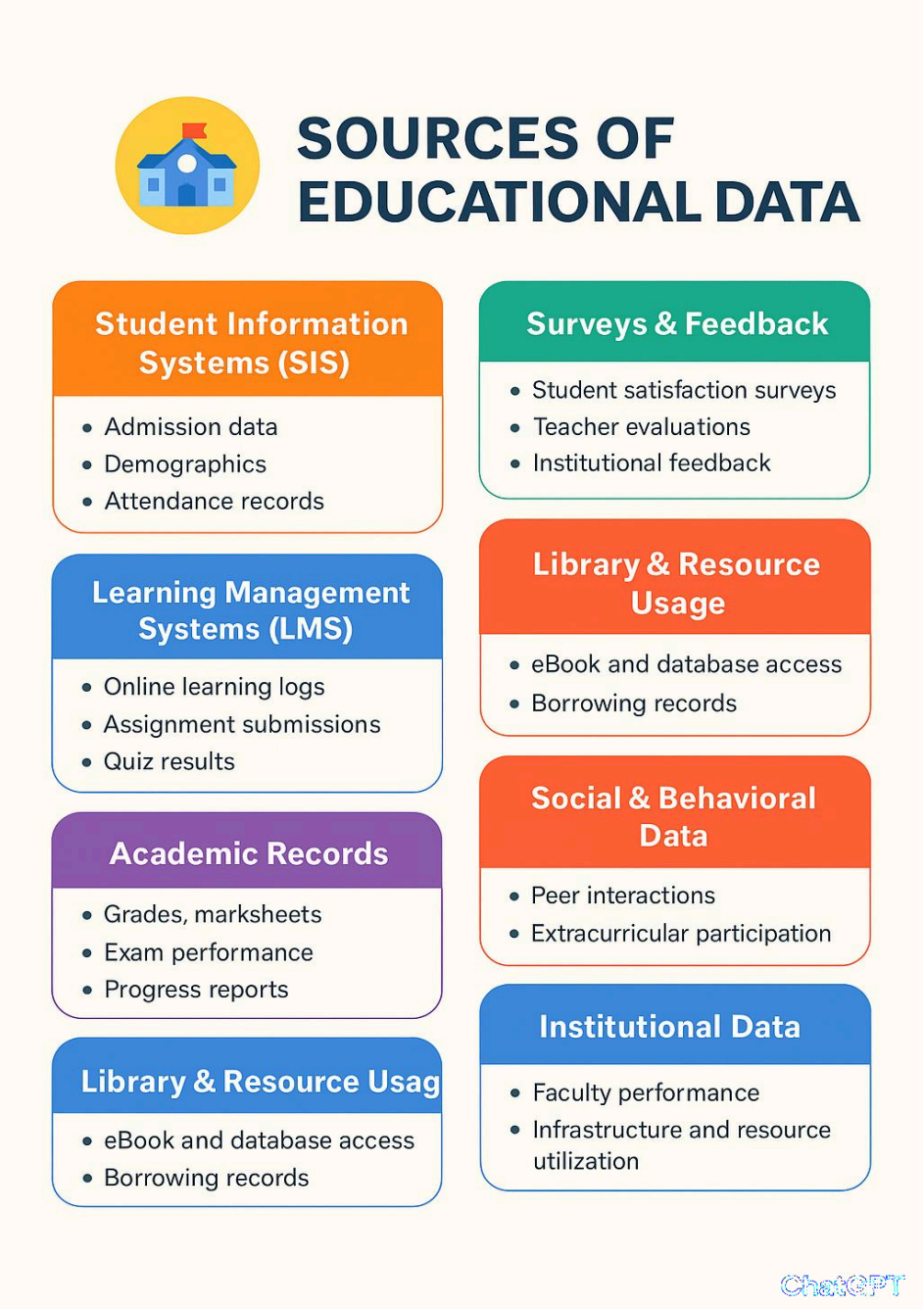
Manufacturing & Supplying

In the context of data mining, manufacturing and supplying, how the raw data is produced (manufactured), processed, and then delivered as useful insights or services. It has huge application uses in quality control & defect detection, predictive maintenance, process optimization, inventory management, and production forecasting. Similarly, application of data mining in. the supply chain includes demand forecasting, supplier selection & evaluation, logistics optimization, fraud & risk detection, and customer order fulfilment. Different types of technology Some key steps in Manufacturing data mining are data generation, data acquisition, data preprocessing, feature engineering, and integration. Similarly, supplying data mining knowledge extraction, visualization & reporting, integration into applications, automated alerts & recommendations, API & Data feeds.



Education

Data mining in education, often called educational data mining (EDM)in this process analyse educational data to uncover patterns that help improve teaching, learning, and administrative decisions. Sources of education include student information systems (SIS) (enrolment, demographics, grades), learning management systems (LMS) (Moodle, Google Classroom, Blackboard logs), online learning platforms (Coursera, Khan Academy interaction data), Assessments & exams (scores, time spent, difficulty levels), surveys & feedback forms (student/ teacher satisfaction), attendance records.



In education, different types of challenges are faced, such as data privacy & ethics (handling sensitive student data responsibly), data quality (incomplete or inconsistent records), and bias in algorithms (risk of unfair predictions). The key data mining techniques commonly applied in educational data mining (EDM), along with their purposes like classification (Decision trees, naïve bayes, support vector machines (SVM), Random Forest), clustering (K-means, Hierarchical clustering, DBSCAN), association rule mining (Apriori, FP-growth), Regression analysis (linear regression, multiple regression, logistic regression), prediction (neural networks, ensemble models), outlier detection/anomaly detection (statistical methods, isolation forest, Z-score, text mining/ natural language processing (NLP) (sentiment analysis, topic modelling (LDA), sequential pattern mining (SPADE, Prefix span).

Finance

In the context of financial data, which is huge, complex, and continuously growing. Data mining involves extracting meaningful patterns, trends, and relationships from financial datasets to support decision-making, detect anomalies, and forecast future trends in various areas, including the banking sector(credit scoring & loan approval, fraud detection in transitions, customer segmentation, ATM cash demand forecasting), stock market & investment(predicting stock prices using time series analysis, risk vs return, identifying market trends & sentiment analysis, algorithmic/ automated trading), insurance, and risk management (Credit risk assessment, market risk prediction, operational risk monitoring, early warning systems for financial crises). It deals with risk management & credit scoring, fraud detection, investment & stock market analysis, customer relationship management (CRM) (personalized financial services, cross-selling & up-selling banking/insurance products, churn prediction, lifetime value prediction), anti-money laundering (AML) (detecting suspicious financial transactions, identifying hidden transaction networks, pattern analysis of fund transfers), insurance & actuarial analysis. To solve these issues, we use classification (credit risk assessment, customer segmentation), clustering (grouping customers with similar spending patterns), association rules (discovering relation between financial products), regression (predicting stock prices, loan repayment), anomaly detection (fraud detection, unusual transactions), time series analysis (forecasting stock trends, currency fluctuations).

Telecommunication

Finance + Data Mining

Banking | Stock Market | Insurance

Credit Risk
Fraud Detects
Customer Segmentation

Price Prediction
Portfolio Opt. Trend Analysis

Fraud Detection
Claim Prediction
Policy Pricing

Risk Management

AML & CRM Applications

Telecommunication companies generate massive volumes of data including call detail records (CDRs), SMS, internet usage, billing, customer support, location, marketing, retail, fraud detection/spam detection, and network management.

Data mining has great applications in marketing (customer segmentation, market basket analysis), customer retention (customer churn prediction), fraud detection (anomaly detection), revenue assurance, service personalization, network management (network fault isolation, resource management) etc.

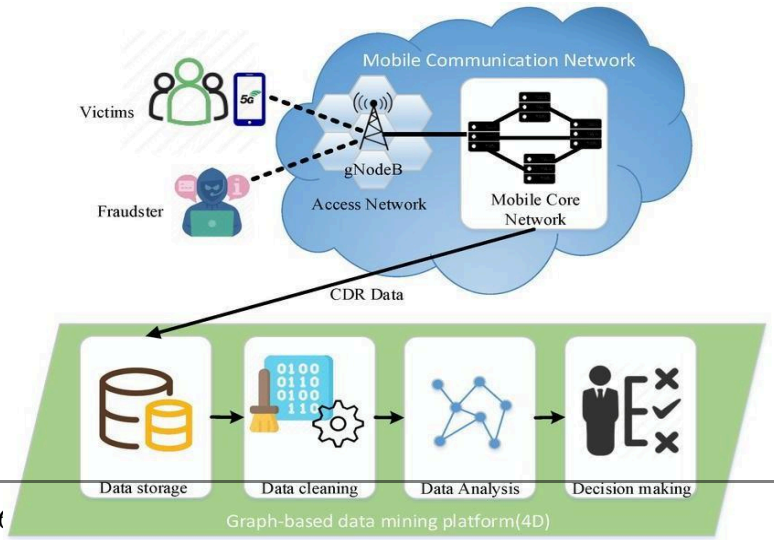
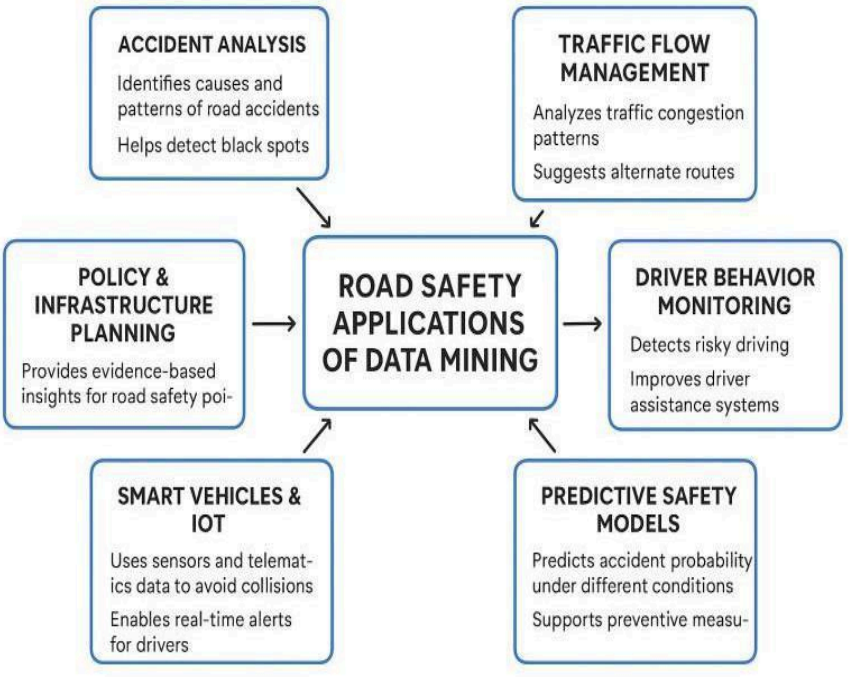


Illustration of fraud detection in mobile

Different types of techniques are used, such as classification (predict churn/fraud), clustering (segment customers), association rule mining (find common usage patterns), anomaly detection (detect network faults/fraud), and time series analysis (forecast traffic load & demand).

Road Safety

This area is considered a huge source of data. Different types of data include accidental data (what happened), road data (where it happened), vehicle data (what was involved), driver behaviour monitoring (who was involved), environmental data (under what conditions), traffic data (flow context), and response data (after-effects). For the analysis, different types of data collection tools are available: traffic sensors & IoT devices (CCTV cameras, GPS trackers, speed sensors, vehicle telematics), mobile & app data (Google maps, Waze, ride-hailing apps), Government databases (police accident reports, hospital injury records), GIS tools (ArcGIS, QGIS for spatial accident mapping). Different statistical packages are SPSS, SAS, R, machine learning platforms (Python, RapidMiner, Weka), clustering & pattern recognition, say K-means, decision trees, and Neural Networks are used to analyse the data. In short by following some steps, the issues can be solved: Data Collection (IoT, GPS, Sensors), storage & processing (databases, Hadoop, Spark), analysis & mining (Python, R, Weka, SPSS), Visualization (Tableau, Power BI, GIS), Simulation (VISSIM, SUMO) etc. Road safety can be described as follows:



Healthcare

Health is not just being free from disease, but a holistic state of well-being that includes physical, emotional, occupational, intellectual, social, financial, environmental, and spiritual aspects, and is not just the absence of disease or infirmity. Health is important, as it increases the quality of life, resilience, and longevity. According to the WHO (1948) Health is a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity." Different slogan-style versions on health used "health is wealth-Protect it every day", "Small steps for you, big gains for health", "Healthy mind, healthy body, and healthy society" etc. The spectrum of health emphasizes that health is a continuous process, ranging from complete wellness to death, highlighting the importance of prevention, early diagnosis, and healthy living. Health is dynamic; people can move forward or backward along the spectrum depending on their awareness, lifestyle, and medical practices. The main components of spectrums are optimal health/ wellness, good health, ill health, disease/disability, and death. Health is closely related to many components of society, because it is both an individual right and a social responsibility.

Health is closely related to many components of society, because it is both an individual right and a social responsibility, like family, education system, economy & employment, culture & religion, government & policy, environment, and healthcare system. A society can progress only when its people are healthy, and in turn, social components must support health.

Different sources of health data:

Individual data/clinical data:

Applications of data mining in healthcare

1. Disease Diagnosis & Prediction, treatment effectiveness,
2. Clinical decision support system,
3. Medical imaging,
4. Healthcare management, fraud detection in medical claims,
5. Genomics and drug discovery.

Data mining techniques used

Here are different methods to solve the issues arising in handling medical data, such as data privacy and security, data Quality, interpretability, and integration. By using these techniques, the result can be

improved with enhanced diagnosis accuracy, better patient outcomes, cost-effective healthcare delivery, efficient hospital management, and improved medical research.

Technique	**Applications**
Classification	Diagnose Diseases [1],[2],[3],[5],[6]
Clustering	Group Patients based on symptoms or risk factors
Association Rule Mining	Find the relationship between diseases and Symptoms
Regression	Predict disease progression or treatment cost
Neural Networks	Analyse complex datasets like ECG or brain scans
Decision Trees	Assist in treatment decision pathways
Support Vector Machines	Prediction tasks, disease diagnosis, prognosis and patient risk stratification [4]

1. **Application of Classification Rule**

Considering the dataset for predicting heart disease in six different patients

Patient ID	Age	Cholesterol	Blood Pressure
1	56	High	High
2	31	Normal	Normal
3	46	High	Normal
4	29	Normal	Normal
5	61	High	High
6	41	Normal	High

Table 1: Patient data

Classification Rule: Binary classification

$$If\ cholesterol = High \rightarrow Heart\ disease = Yes$$

$$Else \rightarrow Heart\ Disease = No$$

Patient ID	Cholesterol	Blood Pressure	Heart Disease(yes/No)
1	High	High	Yes
2	Normal	Normal	No
3	High	Normal	Yes
4	Normal	Normal	No
5	High	High	Yes
6	Normal	High	No

Table 2: Classification of heart disease

In the above data, three people have heart disease, and the remaining are free from this disease.

2. Application of clustering

Here, the study examines clustering (K-Means) of similar patients without predefined categories.

Consider the data of COPD (Chronic Obstructive Pulmonary Disease) patients

Patient ID	Age	BMI $\frac{Weight\ (Kg)}{Height\ (m)^2}$	Smoking (Pack-Year)	Cough (Severity 0-10)	Breathlessness (Score 0-10)	Fever
1	28	22.1	1	1	1	Occasional
2	32	23.4	0	0	2	Occasional
3	29	20.8	3	2	3	Occasional
4	35	24.2	2	2	2	Occasional
5	31	21.7	4	1	1	Occasional
6	27	19.9	1	0	3	Occasional
7	33	22.8	0	1	2	Occasional
8	36	24.4	4	2	1	Occasional
9	30	23.1	2	1	3	Occasional
10	28	20.5	3	2	2	Occasional
11	34	24.6	2	2	1	Occasional
12	29	19.8	1	2	3	Occasional
13	31	22.3	2	1	2	Occasional
14	35	23.9	0	1	1	Occasional
15	33	24.8	3	2	3	Occasional
16	36	21.4	4	1	2	Occasional
17	27	20.7	1	2	1	Occasional
18	38	24.7	3	2	3	Occasional
19	30	22.0	2	1	2	Occasional
20	37	23.3	1	1	2	Occasional
21	45	26.4	20	4	4	Some
22	52	27.1	18	5	6	Some
23	49	28	25	6	5	Some

24	41	25.2	17	5	4	Some
25	55	29.3	22	6	6	Some
26	48	27.5	19	4	5	Some
27	50	28.4	29	5	6	Some
28	53	26.9	20	4	4	Some
29	42	25.7	24	6	5	Some
30	59	29.1	27	5	6	Some
31	44	27.6	18	6	4	Some
32	56	25.9	23	4	5	Some
33	47	28.8	16	5	6	Some
34	51	27.3	19	6	4	Some
35	58	26.2	21	4	5	Some
36	43	25.8	20	5	5	Some
37	49	29.4	28	5	6	Some
38	45	27.1	17	6	4	Some
39	50	28.2	26	4	5	Some
40	54	26.7	22	5	6	Some
41	68	32.8	40	8	7	Common
42	73	35.1	45	7	9	Common
43	70	33.5	41	9	8	Common
44	77	36	43	8	10	Common
45	66	32.9	40	9	7	Common
46	82	35.8	44	10	10	Common
47	71	33.2	42	7	8	Common

48	75	34.6	41	9	9	Common
49	63	31.9	40	8	7	Common
50	79	36.4	44	10	10	Common
51	68	32.5	42	9	7	Common
52	72	35.9	45	7	9	Common
53	81	37.2	43	10	8	Common

54	67	33.7	41	10	10	Common
55	74	36.5	42	8	10	Common
56	76	34.4	44	9	10	Common
57	69	35	43	10	10	Common
58	83	37.9	45	9	10	Common
59	64	32.6	41	10	10	Common
60	78	36.7	44	10	10	Common

Table 3: Patient report

Cluster	Meaning	Feature
Cluster 1	Mild COPD (Young, low BMI, non-smoker, mild symptoms)	$Age < 40$, $BMI < 25$, $Non-Smoker$/ $minimal$, $No\ comorbidity$ / $Occasional$, Cough only/ mild, Breathlessness (1-3), Fever (Occasional)
Cluster 2	Moderate COPD (Middle-aged, moderate smoking & moderate symptoms)	$Age(40-60)$, $BMI(25-30)$, $Smoker(15-30)$, $One\ comorbidity$, Cough Moderate, Breathlessness (4-6) & Some fever
Cluster 3	Severe COPD (Older, heavy smokers, severe symptoms)	$Age > 60$, $BMI > 30$, $Smoker$ $(40-45\ pack-years)$, $Multiple\ comorbidity$ Severe Cough, Fever (Common) & Breathlessness (7-10)

Table 4: Distribution level of cluster

Patient	Age	Smoking	Breathlessness (Score 0-10)	Cough (Severit	BMI $Weight\ (Kg)$	Fever

		(Pack-Year ear)		y 0-10)	$\frac{Weight\ (Kg)}{Height\ (m)^2}$	
1	28	22.1	1	1	1	Occasional
2	32	23.4	0	0	2	Occasional
3	29	20.8	3	2	3	Occasional
4	35	24.2	2	2	2	Occasional
5	31	21.7	4	1	1	Occasional
6	27	19.9	1	0	3	Occasional
7	33	22.8	0	1	2	Occasional
8	36	24.4	4	2	1	Occasional
9	30	23.1	2	1	3	Occasional
10	28	20.5	3	2	2	Occasional
11	34	24.6	2	2	1	Occasional
12	29	19.8	1	2	3	Occasional
13	31	22.3	2	1	2	Occasional
14	35	23.9	0	1	1	Occasional
15	33	24.8	3	2	3	Occasional
16	36	21.4	4	1	2	Occasional
17	27	20.7	1	2	1	Occasional
18	38	24.7	3	2	3	Occasional
19	30	22.0	2	1	2	Occasional
20	37	23.3	1	1	2	Occasional

Table 5: Cluster 1

Patient	Age	Smoking (Pack-Year ear)	Breathlessness (Score 0-10)	Cough (Severity 0-10)	BMI $\frac{Weight\ (Kg)}{Height\ (m)^2}$	Fever
21	45	26.4	20	4	4	Some
22	52	27.1	18	5	6	Some
23	49	28	25	6	5	Some

24	41	25.2	17	5	4	Some
25	55	29.3	22	6	6	Some
26	48	27.5	19	4	5	Some
27	50	28.4	29	5	6	Some
28	53	26.9	20	4	4	Some
29	42	25.7	24	6	5	Some
30	59	29.1	27	5	6	Some
31	44	27.6	18	6	4	Some
32	56	25.9	23	4	5	Some
33	47	28.8	16	5	6	Some
34	51	27.3	19	6	4	Some
35	58	26.2	21	4	5	Some
36	43	25.8	20	5	5	Some
37	49	29.4	28	5	6	Some
38	45	27.1	17	6	4	Some
39	50	28.2	26	4	5	Some
40	54	26.7	22	5	6	Some

Table 6: Cluster 2

Patient	Age	Smoking (Pack-Year)	Breathlessness (Score 0-10)	Cough (Severity 0-10)	BMI $\frac{Weight\ (Kg)}{Height\ (m)^2}$	Fever
41	68	32.8	40	8	7	Common
42	73	35.1	45	7	9	Common
43	70	33.5	41	9	8	Common
44	77	36	43	8	10	Common
45	66	32.9	40	9	7	Common
46	82	35.8	44	10	10	Common
47	71	33.2	42	7	8	Common
48	75	34.6	41	9	9	Common
49	63	31.9	40	8	7	Common

50	79	36.4	44	10	10	Common
51	68	32.5	42	9	7	Common
52	72	35.9	45	7	9	Common
53	81	37.2	43	10	8	Common
54	67	33.7	41	10	10	Common
55	74	36.5	42	8	10	Common
56	76	34.4	44	9	10	Common
57	69	35	43	10	10	Common
58	83	37.9	45	9	10	Common
59	64	32.6	41	10	10	Common
60	78	36.7	44	10	10	Common

Table 7: Cluster 3

2. Consider the study of 100 patients with cardiovascular symptoms

Patient ID	Age	BMI	Blood Pressure	Cholesterol
1	55	28	150	220
2	32	22	120	180
3	60	30	160	240
4	29	21	118	175
5	58	27	155	230

Table 8: Patient data

Patient ID	Age Identification	Cholesterol	Blood Pressure	Risk
1	Older	High	Hypertension Stage 2	High-risk
2	Younger	Normal	Normal	Low-risk
3	Older	High	Hypertension Stage 2	High-risk
4	Younger	Normal	Normal	Low-risk
5	Older	High	Hypertension	High-risk

			Stage 2	

Table 9: Resulting report

Cluster 1: Older, Higher BMI, Higher BP → High-risk group

Cluster 2: Younger, Normal BMI, normal BP → Low-risk group

By clustering, the hidden patterns are found. Three have high risk, and the remaining three have low risk.

3. **Application of Regression Analysis**

Apply multiple linear regression to check how different health factors affect blood pressure (BP)

Patient (Person)	Age (Years)	BMI (Kg/m^2)	Cholesterol (mg/dl)	Exercise (Hr)	Smoking (No. of cigarettes/day)	Blood Pressure (BP)
1	25	22	180	6	0	110
2	32	25	190	4	5	118
3	45	28	210	3	10	132
4	36	24	195	5	2	120
5	50	30	230	2	15	145
6	41	27	200	4	8	128
7	55	31	240	1	20	150
8	29	23	185	5	3	115
9	47	29	220	2	12	140
10	60	32	250	1	18	155

Table 10: Patient data

Consider a linear model

$$BP = \beta 0 + \beta 1(Age) + \beta 2(BMI) + \beta 3(Cholesterol) + \beta 4(Exercise) + \beta 5(Smoking) + \in$$

By using the software found the coefficient

Coefficient	$\beta 0$	$\beta 1$	$\beta 2$	$\beta 3$	$\beta 4$	$\beta 5$
Value	60	0.4	1.5	0.1	-1.8	0.9

Table 11: Effect report

Higher age, BMI, cholesterol, and smoking are positively associated with blood pressure and more exercise is negatively associated (reduced BP).

4. **Application of the decision tree**

Consider the data

Patient	Age(yr)	BMI (kg/m^2)	Fasting blood sugar (mg/dL)
1	35	24.5	95
2	42	28.1	130
3	51	26.7	145
4	29	22.4	85
5	48	31.5	160
6	39	27.3	120
7	55	29.2	140
8	33	23.8	98
9	60	30.1	170
10	45	25.6	110

Table 12: Patient data

Step 1:

Diabetic	$FBS > 125\ mg/dL$
Not Diabetic	$FBS \leq 125\ mg/dL$

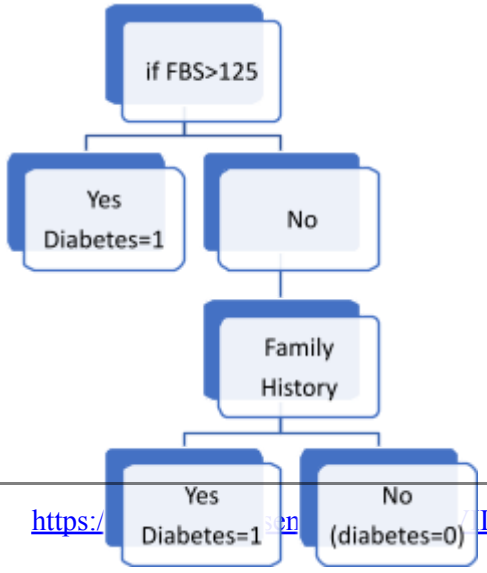
Table 13: Fasting Blood sugar (FBS)

Step 2:

Family history	Diabetes
Yes →	Yes
No →	No

Table 14: Family history v/s diabetes

Step 3: **Decision tree**



5. Application of association rule mining

Let us consider the hypothetical data to see the association between factors for gallstone surgery level.

Variable items used in ARM

Symptoms	Investigations	Risk factors
RUQ pain	Elevated WBC	Age>50
Fever	Elevated bilirubin	Female
Vomiting	Raised ALP	Obesity
Jaundice	CBD stone on USG	Diabetes
Pain after fatty meals	Gallbladder wall thickening	
	Pericholecystic Fluid	

Table 15: Distribution of factors

Used the variables those highly correlated to study

ID	Fever	WBC↑	RUQ Pain	Jaundice	GB Wall Thick	CBD Dilated	Stones on USG
1	Yes	Yes	Yes	No	Yes	No	Yes
2	No	No	Yes	No	No	No	Yes
3	No	No	No	No	No	No	Yes
4	Yes	Yes	Yes	Yes	Yes	Yes	Yes
5	No	No	Yes	No	No	No	Yes
6	No	No	No	No	No	No	Yes
7	Yes	Yes	Yes	No	Yes	No	Yes
8	No	No	Yes	No	No	No	Yes
9	Yes	Yes	Yes	Yes	Yes	Yes	Yes
10	No	No	No	No	No	No	Yes
11	No	No	Yes	No	No	No	Yes
12	Yes	Yes	Yes	No	Yes	No	Yes
13	No	No	No	No	No	No	Yes
14	Yes	Yes	Yes	Yes	Yes	Yes	Yes
15	No	No	Yes	No	No	No	Yes

Table 16: Patient data

State	Condition	Patients (Antecedent)	Outcome	Patient (Consequent)
1	Fever +High WBC + GB Wall thickening & Jaundice + CBD dilation →	1, 4, 7, 7 ,9, 12, 14	Emergency surgery →	1, 4, 7, 7 ,9, 12, 14
2	RUQ Pain + gallstones on ultrasound →	1, 2, 4, 5, 7, 8, 9, 11, 12, 14, 15	Elective surgery →	2, 5, 8 11, 15
3	Gallstones + normal liver function test →	3, 5, 6, 8, 10, 11, 13, 15	No surgery →	3, 6, 10, 13

Table 17: Applicable Conditions

State	Support	Confidence	Lift	Structure
1	0.40	1.00	2.50	Complicated stones
2	0.733	0.455	1.36	Symptomatic stones
3	0.533	0.50	1.87	Asymptomatic stones

Table 18: Resulting reports

It is just tried to understand the patterns match real-world clinical behaviour by taking ARM results show that specific combinations of symptoms, lab values, and ultra sound finding s strongly predict whether gallstone patients require emergency surgery, elective surgery, or no surgery.

6. Application of support vector machines

(A) Let us consider an example of diabetes to understand the message conveyed by the coefficient of the linear SVM model

Patient	BMI $X1$	FBS$X2$
1	22	85
2	28	140
3	31	160
4	24	95
5	29	155
6	21	90

Table 19: Patient

data In the linear SVM, the decision function is

$$(x) = \omega 1(BMI) + \omega 2(FBS) + b$$

$\omega 1, \omega 2\ are\ coefficients, \omega 1\ represents\ effect\ of\ BMI\ and\ \omega 2\ represents\ effect\ of\ BMI.$

$b\ is\ bias\ intercept, and\ it\ shifs\ the\ decision\ boundary.$

After applying a linear SVM, one reasonable separating hyperplane could be

$$(x) = 0.6(BMI) + 0.08(FBS) - 25$$

Patient	BMI $X1$	FBS$X2$	Diabetes (Y)
1	22	85	-1
2	28	140	+1
3	31	160	+1
4	24	95	-1
5	29	155	+1
6	21	90	-1

Table 20: Resulting report about diabetes

Here ω_1, It is positive, so there is a BMI increase in diabetes risk, or say BMI contributes moderately and ω_2, it is a large, stronger influence than BMI or say FBS has a strong influence. The prediction rule is $(x) > 0$ The patient is a diabetic if $f(x) < 0$ then the patient does not have diabetes.

Conclusion

In this paper different types of uses of data mining is describe specifically medical uses shown. The maturity of medical data describes several challenges, like issues of privacy and security, data quality, interpretability, and integration across healthcare systems. Different types of techniques, such as analytical and machine learning, can be used to address these challenges and enhance healthcare outcomes. The easiest method for classification supports disease diagnosis, while clustering helps in grouping patients based on symptoms or risk factors. Association rule mining identifies hidden relationships between diseases and symptoms, and regression techniques predict disease development, worsening, or spreading over time inside the body or treatment costs. Neural networks enable the analysis of complex medical datasets such as ECG signals and brain imaging. Decision tree provide decision-making, whereas support vector machines contribute to accurate prediction, diagnosis, and patient risk stratification. By employing these methods, healthcare systems can achieve improved diagnostic accuracy, enhanced patient outcomes, cost-effective treatments, efficient hospital management, and strengthened medical research.

In the context of disease diagnosis, binary classification techniques are often used. This method categorises patients into two possible outcomes, such as the presence or absence of heart disease. For example, using patient data such as age, cholesterol level, and blood pressure, a binary classification model can help predict whether a patient is at high risk of developing heart disease. This assists doctors in early detection and timely treatment.

Additionally, clustering methods are applied to group patients with similar characteristics. Unlike classification, clustering does not rely on predefined categories; instead, it identifies hidden patterns within the dataset. For instance, clustering might reveal subgroups of COPD patients who share similar symptoms, age, smoking, body mass index, breathlessness, cough, and risk factors. Such insights help tailor treatment plans and understand disease progression more effectively.

Furthermore, regression analysis is used when it is necessary to examine the relationship between a dependent variable (such as blood pressure) and one or more independent variables (like age, BMI, cholesterol, exercise, and smoking habits). This helps in predicting medical outcomes numerically. Regression can predict how much a patient's risk of blood pressure increases with a rise in age, BMI, cholesterol, exercise, smoking, etc.

The support vector machine method is a powerful predictive tool for identifying heart disease in patients using non-invasive clinical measurements. That is one of the effective classification performances of heart disease.

Overall, these data mining techniques—binary classification, clustering, and regression—work together to improve medical diagnosis, enhance understanding of disease patterns, and support evidence-based decision-making in healthcare.

References

- Deulkar, D.S. and Deshmukh, R. R. (2016). Data Mining Classification. Imperial Journal of Interdisciplinary Research, 2
- Ekwonwune, E., Ubochi, C. and Duroha, A. (2022). Data Mining as a Technique for Healthcare Approaches. International Journal of Communications, Network and System Sciences 15, 9, 149-165.
- Kamma, S., Parul, B., Sandeep, D. and Sudluv (2016). Analysis of Application of Data Mining Techniques in Healthcare. International Journal of Computer Applications, 148, 16-21.
- Ozlem Akay and Mehmet Tunceli (2021). Use of the Support Vector Regression in Medical Data Analysis. Experimental and Applied Medical Science, 2, 4, 242-256.
- Poornima, S., Sanjay, S., Gayatri, S. (2018). Effective heart disease prediction system using data mining techniques. International Journal of Nano medicine, 121- 124.
- Srinivas, K., Rani, K. and Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, 2, 250-255.