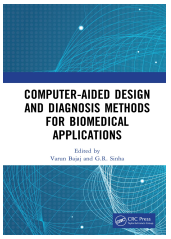


Chapter

Improved Classification Techniques for the Diagnosis and Prognosis of Cancer

By Pankaj Dadheech, Ankit Kumar, S. R. Dogiwal, Vipin Jain, Vijander Singh, Linesh Raja

Book Computer-aided Design and Diagnosis Methods for Biomedical Applications (https://www.taylorfrancis.com/books/mono/10.1201/9781003121152/computer-aided-design-diagnosis-methods-biomedical-applications?refId=29d52d05-f616-4c4f-b8b4-ecdfd5d3abde)

Edition	1st Edition
First Published	2021
Imprint	CRC Press
Pages	24
eBook ISBN	9781003121152

ABSTRACT

(/)

12 Improved Classification Techniques for the Diagnosis and Prognosis of Cancer

Pankaj Dadheech, Ankit Kumar, S. R. Dogiwal, Vipin Jain, Vijander Singh, and Linesh Raja

CONTENTS

12.1 INTRODUCTION

The healthcare market in India is one of the largest and fastest-growing industries in the world, it consumes nearly 10% of the GDP of a developed or developing nation; the healthcare industry contributes a significant amount to the country's economy. The Indian healthcare sector provides new and existing players with special opportunities for achieving and performing innovative research. Healthcare in India was also awarded "polio-free" status by the World Health Organization (WHO). According to research by McKinsey & Company, in the next decennary, consumer awareness and demand for better services and facilities will increase, and in India the healthcare industry will become the third-largest service sector employer. The latest innovation in healthcare data mining is the "big data analytics" revolution. In the healthcare industry, big data consists of electronic health datasets or flat-file data which are disordered, complex, and so large that they are nearly impossible to manage with the available tools or traditional hardware and software techniques. For the healthcare data/information, there is a very large amount of data available for understanding the patterns and trends; hence, big data analytics has the potential to improve healthcare services and provide cost reductions. This chapter explores data mining applications, and shows the difference these can make to the patients and their daily lives; it will also provide suggestions for future work/directions in healthcare. The hospital-based survey also provides benefits for various data mining techniques, as it can show results in different ways, such as clustering, association rules, and classification in the healthcare domain. This chapter also defines cancer and morphology patterns

among various patients in Haryana and the surrounding state with the help of the above-defined different data mining techniques [1].

12.1.1 Medical Services in India

If the Indian economy grows faster than the economies of developed nations and the education rate keeps on increasing, then much of the Indian will be middle class by 2025, and the middle class can afford quality healthcare. According to a CII study, India needs 50 billion dollars annually to fulfill its healthcare requirement for the next 15 years, until 2040; India needs 2 million beds, and requires an immediate investment of 82 billion dollars. According to the PWC, 60% of patients are outpatients in the private sector. Nearly 40% of hospital beds are in the private sector. Around 30% of the medical market is covered by this economic segment. Now, in the Indian market, hospitals are realizing that IT can be effective and efficient for hospital growth. Indian healthcare services are fast-growing according to a CII-McKinsey study; Indian hospitals are the first choice of foreign tourists for health diagnosis. India is gaining a significant reputation for medical tourism from Gulf countries. Figure 12.1 shows the various shares of healthcare spending in India. Data mining techniques are used to diagnose different cancers, such as the early detection of breast cancer, which is one of the leading cancers in Asian countries.

12.1.2 Data Mining in Field of HealthCare

Data mining explores the hidden patterns or information in a data warehouse. This knowledge, i.e., information extracted from a vast dataset, is presented in an understandable form. Later on, HMIS (Healthcare Management Information System) is in the healthcare domain and fake cases are found. Healthcare and commercial databases are growing at an unpredicted rate. To handle these huge data sets, we need mature data mining algorithms which can be combined with older statistical methods. Table 12.1 shows the data to knowledge evolution.

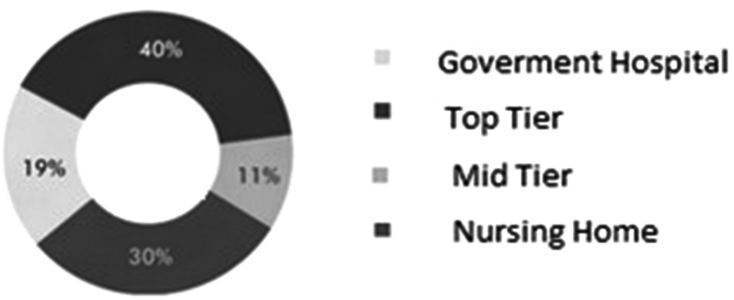


FIGURE 12.1 Shares in healthcare spending in India.

TABLE 12.1
Data to Knowledge Evolution

Evolutionary Step Year Wise	Questions From Business perspective	Technologies Used	Product Providers
Data Collection (1960–1980)	"What was the total revenue for the hospital in the last five years"	Computers, tapes, disks	CDC, IBM
Accessing Data (1980–1990)	"Total number of patient available department wise in the hospital"	Structured query language (SQL), relational databases (RDBMS)	IBM, Microsoft, Informix, Oracle, Sybase
Data Warehousing and Decision Support (1990s)	"Total number of patient available department wise in the hospital. Drill down to a single patient"	Data warehouses, multi-dimensional databases, OLAP	Micro strategy, Arbor, Cognos, Comshare, Pilot
Data Mining (Emerging Technology)	"How many patients will require cosmetics this month? Why?"	Massive databases, advanced algorithms, multiprocessor computers	SGI, IBM, Pilot startups, Lockheed

12.1.3 Architecture for Data Mining

Data mining tools and techniques operate by extracting, importing, and analyzing the data. A hospital data warehouse contains a combination of patient details and hospital details, i.e., patient beds and medicines. This warehouse information can be collected in variety of relational database systems, such as Sybase, Oracle, and MySQL, in an optimized manner so that it can be easy and fast to access. Now the OLAP (online analytical processing) database comes into the picture. With the help of facts and dimensions, a multi-dimensional structure is created, which helps users to analyze data for business purposes. The data mining techniques must be merged with the data warehouse and the OLAP server to produce new predictions and results.

Figure 12.2 shows the knowledge discovery process.

i. **Data cleaning** removes the inconsistent data.
ii. **Data integration** combines data from various sources.
iii. **Data selection and transformation** extracts the relevant data which is used for further analytics.
iv. **Data mining** contains intelligent methods (algorithms) to extract patterns out of the data (pattern evolution).
v. **Knowledge discovery** contains the overall visualization of the data.

12.1.4 Data Mining in Healthcare

Generally, healthcare data is available in flat-files, relational databases, or in advanced database systems, such as images which are collected from different data sources,

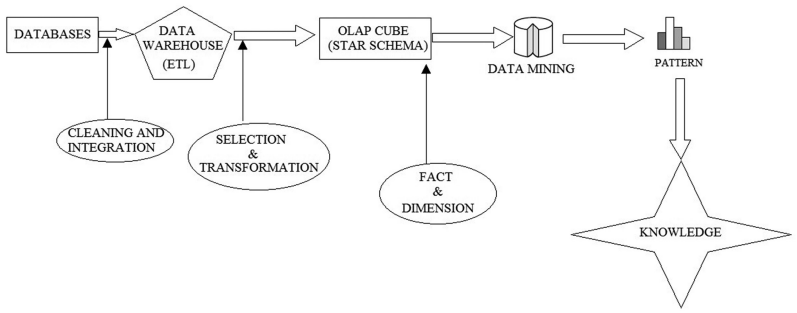


FIGURE 12.2 Knowledge discovery process.

including OPD, laboratory information systems, operation theater modules, radiotherapy modules, chemotherapy modules, blood banks, and the drug store. Various types of healthcare data are represented in Table 12.2.

12.1.4.1 Nature of Healthcare Data

Healthcare data is very specific. To mine healthcare data, all the information needs to be changed into numeric values. The methods for this task are described in medical textbooks which usually come from a certain range, and it is beyond the scope of this chapter. With the help of numeric data it is easy to apply mining algorithms to extract knowledge out of the data.

12.1.4.2 Patient Data Set

Data mining determines knowledge from a patient's dataset, and this dataset is a collection of different data sources, as mentioned above. In this research work, we collected data from the RCCR (Regional Cancer Center Rohtak) and prepared and analyzed the datasets for the different regions of Haryana and its surrounding states; then we applied a data mining algorithm to extract hidden patterns. This helped us to determine the cancer site and morphological patterns of various patients, and it also helped to develop a suggestive management information system to improve cancer treatment.

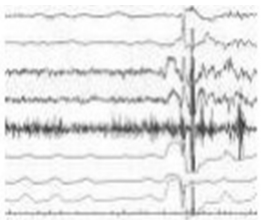
12.1.4.3 Preliminary Analysis of Dataset

Preliminary analysis of a dataset is an essential step for transforming unstructured data or row data into a format suitable for applying data mining techniques and improving the quality of the data. Preliminary analysis is to identify hospital and government needs, perform economic, technical analysis and green analytics, perform a cost–benefit analysis, and create a suggestive management information system. There should be enough expertise available for database queries and data mining algorithms and software for doing the analysis.

12.1.5 Medical Data Selection and Preparation

Medical data selection and preparation (MDP) is a crucial and very time-consuming process for data mining. It takes around 45% to 55% of the total time

TABLE 12.2
Types of Healthcare Data

ECG	
EEG	
RTG	

to prepare a dataset for the data mining process. Healthcare data selection and processing aims to establish a data warehouse for data mining algorithms and information sharing. MDSP is connected to various data sources, such as laboratories, operation theaters, blood banks, drug stores, and therapy modules, which hold valuable data, such as patient information, including their workflow and past prescriptions. There are two processes involved in MDSP, namely data selection and the processing. The medical data selection process selects data from different sources, as mentioned above. And all the relevant data is updated at a centralized data warehouse according to the patient details. Then, the MDS process selects whether the data is useful or not and then passes it to the research department. Medical data preparation involves examining or logging the data and verifying the data for accuracy and exactness. This is because some selected data may be missing or in different formats. At this stage, all necessary information is extracted from the data selection process to apply to further data mining processes, the age of the patient should be "81" but recorded as "18" so this comes under the human error. Data cleaning can reduce the missing, inconsistent, and noisy data that affect the results.

12.1.6 Issues and Challenges

i. **Domain expertise:**
   Domain knowledge related to healthcare data is very important; it helps in finding out different patterns from the database.
ii. **Visualization of healthcare data mining result:**
   After discovering the knowledge, the results should be easily understood and directly usable by humans.
iii. **Handling incomplete, noisy, and diverse data:**
   Healthcare data contains exceptional cases and incomplete data objects. Therefore the accuracy of the discovered patterns will be low. For exceptional cases, we have to apply data cleaning and data analysis methods.
iv. **Pattern discovery:**
   It is very difficult to cover thousands of patterns; hence, many patterns are undiscovered, and many of the patterns are uninteresting to the user.

12.1.7 Cancer Treatments Using Decision Support System

To design a DSS, the two main factors involve stakeholder involvement and the type of decision they want to make. A healthcare DSS should be well organized, including procedures for decision-making, strategic planning, and structures according to the government regulations should be kept in mind. The effectiveness of a DSS depends on the methodology used to design the system. We can divide this into three approaches:

i. Clinical algorithms
ii. Heuristics approaches
iii. Mathematical approaches

Data scientists have revealed that neural network systems can provide good planning for patient care, length of patient stay, and mortality rate. However, to implement a DSS with this methodology, extensive research and resources are required. Another very popular approach is data mining techniques, which help to identify rules and patterns concerning various problems. Data mining DSS is built based on the data, and it is effective for cost reduction and improving quality of care. As the nurses and doctors have to deal with various complex diagnoses, it becomes time consuming for them to adopt a new technology or a system. That is why DSSs are not widely accepted, due to time complexity and other constraints, but they are perceived as effective and efficient to use. Figure 12.3 shows the simulation model of a decision support system and Table 12.3 shows the various uses of data mining algorithms.

12.2 REVIEW OF LITERATURE

In the last section, we discussed various data mining fields, including medical services in India, data mining in a healthcare context, issues and challenges in

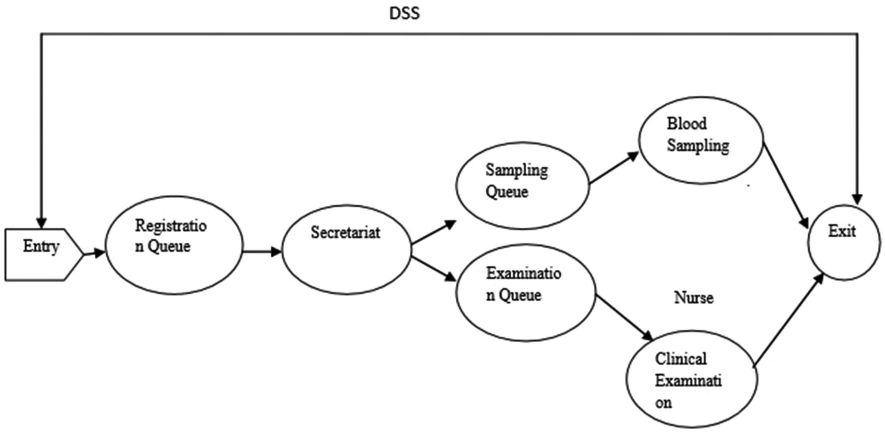


FIGURE 12.3 Simulation model of DSS.

TABLE 12.3
Uses of Data Mining Algorithm

Example of Task	Algorithm to Use
Predicting a discrete value	Naïve Bayes algorithm,
Categorization of a patient according to outcomes and explores related factors.	Decision trees algorithm, Neural network algorithm,
Calculating the probability of server failure within the next four months.	Clustering algorithm.
Predicting an attribute	Linear regression algorithm,
Predict next year sales.	Time-series algorithm,
Calculating or generating a risk factor or score on a given demographics.	Decision trees algorithm.
Sequence prediction	Sequence clustering algorithm.
Capturing and analyzing the sequential activities of outpatient visit, to formulate common activities.	
Searching similar items in transactions:	Association algorithm,
Use of market basket analysis.	Decision trees algorithm.
Suggesting additional product to buyer for purchase.	

healthcare data mining, and data mining applications in cancer treatments. Now, we will discuss research works on data mining in the healthcare industry. In the first section of the chapter, the review process is described. This covers a range of papers that I have studied and about how I conducted my review. The next section consists of a summary of all the papers that I researched to perform the current study and achieve the objective. This section is further divided into various categories based on the solutions in the field of our objective. The third section covers solutions given by various researchers. Then in Section 4, there is a discussion on the strengths and weaknesses which I have observed after studying various papers. The last section covers a summary of all the papers.

12.2.1 Review Process Adapted

Data mining is a term that has been used since the last century in various aspects. We see a new way of using data mining for the benefit of the humanities. If we continue to research data mining, it'll take years. But in the healthcare field, it is a newly emerging area of research. Many healthcare centers are now taking advantage of data mining applications. To perform this research work, we adopted a hierarchical concept of generalization to specialization. After reviewing the healthcare analytics of the world, from the country to the state, we came to Haryana to start experiments with the healthcare data in Haryana and its surroundings.

In this review process, all the research papers were classified based on their solutions.

We reviewed three main categories:

i. **Algorithms used in data mining:** This category involved the review of those papers which were based on the various algorithms used in the data mining process; how researchers used those algorithms and what solutions were provided by them.
ii. **Cancer causes and treatments:** This category involved papers based on cancer, its causes, symptoms, prevention, and cures, which are used by doctors across the world. It also included papers on the various solutions used by medical scientists to cure cancer patients.
iii. **Data mining in the field of healthcare:** This category consisted of review papers based on the application of data mining in the healthcare domain. This showed the devices, therapies, and procedures used by doctors and the works of medical scientists in the development of new equipment to help prevent this disease.

12.2.2 Categorical Review of Literature

12.2.2.1 Literature Review on Algorithm Classification

In 2012 Patil [2] launched an advanced wireless sensor network that aimed to provide online health forecasts through the real-time monitoring of critical body signals. For integrating and executing historical patient data, they implemented cluster algorithms (graph theoretical, *k*-means). The comparative tests of vital signals from clustered algorithms added additional measurements to the hazard warnings and made it more accurate for the doctor to diagnose.

A regression model was created by Carel et al. [3] for asthma drug use using a KDD approach to time-series datasets for historical asthma medication. The clustering and decision tree algorithms were used on the geographic patient sample. The results showed that 274 asthma patients received 9319 approved drugs; the classification also showed that corticosteroid drugs were the most significant indicator of the trend.

In 2014 Reyes et al. [4] were working to build an integrated method for the study of primary healthcare, using clustering methods (partitioned algorithms).

They used Java 12.6 as a language Eclipse 3.4 and JBoss 4.2 as a server to build the solution.

In 1999, Goil et al. [5] discussed a multi-dimensional device scalability framework for OLAP as well as OLAP data mining integration. It generated massive datasets on parallel computers with distributed memory.

12.2.2.2 Literature Review on Cancer Causes and Treatments

In 2013 Kawsar Ahmed et al. [6] gathered data of 400 cancer and non-cancer patients from numerous diagnostic centers, and applied a *k*-means clustering algorithm for the detection of important and irrelevant results. Finally, they established a test for lung cancer that was very effective in identifying a person's lung cancer predisposition.

In 2013 they also discussed the use of classification based technologies, such as artificial neural network guidelines, naïve Bayes, and decision trees, concerning healthcare results (also see Krishnaiah et al. [7]). They focused on common pulmonary symptoms, such as weight loss, pain in the legs and arms or chest, and shortsightedness. They introduced an early warning approach to help save lives.

12.2.2.3 Literature Review on Data Mining in Health Care

In the year 2013, a novel method of data mining was implemented by Akay, Dragomir, and Erlandsson [8], which tracked the experience of diabetic patients with drugs and medical devices. The paper explained how forums were turned into the vectors of search patterns in response to the patients' feedback on their prescriptions and computers. It also offered the impression it could be success for opioid patients.

Subhashet al. in 2013 [9] researched the prevention of infant hunger in developed countries using data processing methods and strategies. They included a selection of literature and used a decision tree methodology. They concluded that information obtained from surveys could be used to more efficiently mitigate infant malnutrition and could also be utilized for potential forecasts.

In 2012 Durairaj, Sivagowry, and Persia [10] addressed methods of data processing for the useful compilation of information from heart disease treatment systems. The study contrasted the efficiency of decision-making, naïve Bayes, neural network, and *k*-mean strategies for cardiac diagnosis extraction. The tools that are used for grouping, clustering, and membership purposes. Like text mining, medical data prediction can be further strengthened.

In 2009, Bellazzi [11] suggested providing remote medical centers with clinical decision support systems (CDSS). This artificial intelligence platform develops frameworks focused on information and data processing. It also defines FOCL hybrid algorithms which are used in the decision support framework and provide a highly efficient, modified version of FOCL.

In 2012, Xylogiannopoulos et al. [12] introduced a middleware model internet inter-orb protocol (IIOP) which was used in a distributed system and resulted in efficient response times between client and server. This model could be used in health clinics for cost reduction and performance efficiency. It could further be applied in medical referral systems and electronic consultation systems.

In 2010, Santhi et al. [13] suggested methodological problems for assessment of the data mining model, translation bioinformatics and bioinformatics aspects of genetic epidemiology on data collection, and data-driven approaches in the field of medical computing, data aggregation, and convergence.

In 2010, Sung Ho Ha et al. [14] stressed the current healthcare issues and their applications; they discussed how healthcare data mining applications were growing in number and producing better healthcare services and policy, and detecting diseases and preventing deaths in hospitals. It also discussed fraudulent insurance claims. This paper gave a broad idea of how to extract knowledge from a database.

12.2.2.4 Issue Wise Solution Approach

Table 12.4 shows various issues with solution approaches used in data mining.

12.3 PROBLEM STATEMENT AND OBJECTIVES

12.3.1 Problem Statement

"Improvisation of Data Mining Techniques in Cancer Site among Various Patients using Market Basket Analysis Algorithm."

Cancer is one of the most deadly diseases worldwide, and many people are currently suffering from this disease. There is no such treatment that provides 100% successful cure of this disease. With the increase of technologies in this area, doctors can explore this disease more and more. One of the best fields of IT, i.e., data mining, is showing very good results in the healthcare domain.

Data mining mainly includes three phases that help to find patterns among database artifacts and examine cancer sites. We study data or medical records by collecting them from any number of hospitals and then analyze that data to find the patients suffering from cancer. There are many types of cancer. From these datasets, we can extract the type of cancer, no of patients suffering from each type of cancer, and the precautions and medicines provided to them in Haryana and its surrounding areas.

To examine this, software vendors were developed for integration, analysis, and reporting services. We used Microsoft tools to generate reports for cancer sites. This study could be continued further and would be helpful for curing large numbers of cancer patients.

12.3.2 Objectives

Four objectives were set up to critically analyze patient data, and the objectives of the research work were as follows:

a. To study the successful data mining techniques and tools to improve the diagnosis of health diseases.
b. To study the various types of algorithms in data mining.
c. To study healthcare data and discover patterns to develop a suggestive management information system to improve cancer treatment.
d. Reducing the mortality rate due to cancer and increasing health awareness.

TABLE 12.4
Issue Wise Solution Approaches Used in Data Mining

Author	Publication Year	Approaches	Accuracy
Yan et al. [15]	2003	Multilayer perceptron	63.61%
Andreeva, P. [16]	2006	Kernel density	84.45%
		Neural network	82.78%
		Decision tree	75.71%
		Naïve Baye	78.58%
Hara et al. [17]	2008	Immune multi-agent neural network	82.33%
		Automatically defined groups	67.81%
Sitar-Taut et al. [18]	2009	Decision tree	60.42%
		Naïve Bayes	62.13%
Chang et al. [19]	2009	Decision tree with sensitivity analysis	86.88%
		Decision tree	90.86%
		Artificial neural network	92.61%
Rajkumar et al. [20]	2010	*k*-NN	45.67%
		Decision tree	52.04%
		Naïve Bayes	52.33%
Srinivas et al. [21]	2010	One dependency augmented Naïve Bayes classifier	80.46%
		Naïve Bayes	84.14%
Anbarasi et al. [22]	2010	Genetic with classification via clustering	88.35%
		Genetic with naïve Bayes	96.53%
		Genetic with decision tree	99.21%
Kangwanariyakul et al. [23]	2010	Bayesian neural network	78.41%
		RBF-kernel support vector machine	60.76%
		Polynomial support vector machine	70.55%
		Linear support vector machine	74.53%
		Probabilistic neural network	70.57%
		Back-propagation neural network	78.42%
Osareh et al. [24]	2010	SVM-POLY	95.29%
		PNN	92.83%
		SVM-RBF	95.44%
		k-NN	94.16%
Abdi et al. [25]	2013	AR_MLP	97.25%
		AR_PSO-SVM	98.93%
		SVM	94.57%

12.3.3 Tools and Technologies Used

12.3.3.1 SQL Server Integration Services (SSIS)

SQL server integration services developed by Microsoft are tools or platforms for performing ETL operations. Table 12.5 shows the features of the SQL server integration services, and Table 12.6 shows the features of a data warehouse.

TABLE 12.5
Features of SSIS

Feature	Datacenter	Standard	Enterprise
Import and export wizard for SQL server	Yes	Yes	Yes
In-build data source connector	Yes	Yes	Yes
Runtime and designer for run time	Yes	Yes	Yes
Task used by import and export wizard	Yes	Yes	Yes
Logging and log provider	Yes	Yes	Yes
Data profiling tools	Yes	Yes	Yes
Extensibility of programmable object	Yes	Yes	Yes

TABLE 12.6
Features of a Data Warehouse

Feature	Datacenter	Standard	Enterprise
Data warehousing and auto-staging	Yes	Yes	Yes
Change in captured data	Yes	Yes	—
Compression in data	Yes	Yes	—
Query optimization with the help of star join	Yes	Yes	—
Automatic view by on query optimizer	Yes	Yes	—
Cubes partition	Yes	Yes	—

12.3.3.2 SQL Server Analysis Services (SSAS)

Table 12.7 shows the features of SQL server analysis services, and Table 12.8 shows the data mining features in SQL server analysis services.

12.4 METHODOLOGIES/ALGORITHMS USED

12.4.1 Naïve Bayes Algorithm

The naïve Bayes is a Bayes-based classification algorithm. It is used to model forecasts. Quite often, we use this algorithm to rapidly produce a mining model to find relationships between inputs and predetermined columns [26]. This model can be used for initial data scans. The naïve Bayes algorithm output screen can be seen in Figure 12.4.

Data required for naïve Bayes models:

- Key column: Each model should contain one primary column.
- Input columns: Input columns must be either discrete or continuous, and they should be independent of each other.
- Predictable column: There should be one predictable attribute which must contain continuous or discrete values.

TABLE 12.7
Features of SSAS

Feature	Datacenter	Standard	Enterprise
Backup facility	Yes	Yes	Yes
Dimension and cube design	Yes	Yes	Yes
Power Pivot	Yes	Yes	Yes
Distributed and Partition cubes	Yes	Yes	—

TABLE 12.8
Data Mining Features in SSAS

Feature	Datacenter	Standard	Enterprise
Bunch of comprehensive data mining algorithm	Yes	Yes	Yes
Integrated tools: Editors, model query builder, wizards, viewers	Yes	Yes	—
Tuning optimization for algorithm	Yes	Yes	—
Pipeline and text mining	Yes	Yes	—
Sequence prediction	Yes	Yes	—

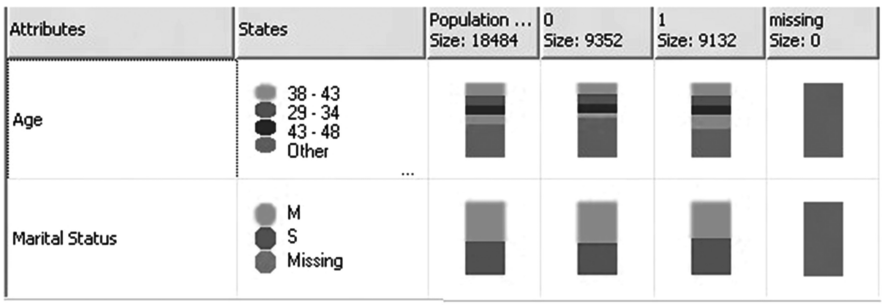


FIGURE 12.4 Naïve Bayes algorithm output screen.

12.4.2 Clustering Algorithm

A clustering algorithm is a segmentation algorithm which uses iterative techniques to make clusters which contain similar type of characteristics. These clusters are useful for finding similar objects in the data for prediction. Figure 12.5 depicts how these clusters look.

A clustering algorithm primarily identifies the relations in the dataset and generates clusters based on that relationship. We can visualize the grouped data with a scatter plotter. Each scatter plot represents cases in a data set. After defining the

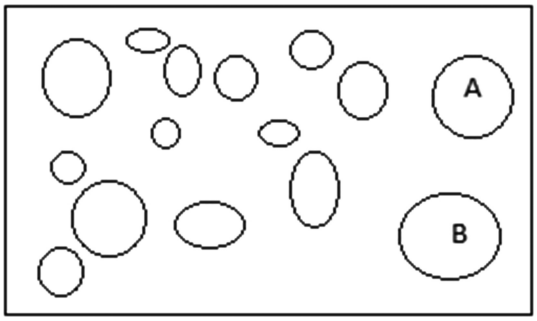


FIGURE 12.5 How cluster looks like.

cluster, it will again calculate how well the cluster represents the grouping of the point, and it will redefine the cluster to better represent the data [27].

Data required for a clustering algorithm:

- A single key column: The model should contain the primary key.
- Input columns: Each model should contain one input column.
- Predictable column: This column is an optional field in the model.

12.4.3 Neural Network Algorithm

In the mining, model networks are dependent on the number of states (input columns and predictable columns) present in them. The neural network algorithm contains three layers of neurons in a network which is created by this algorithm. These three layers contain input–output layers and an optional hidden layer [28].

The input layer determines all the values of the input properties and the data mining process probabilities.

The hidden layer receives feedback from the neuron origin and produces neuron output. This layer includes the weights for the different inputs. The input to the hidden neuron is defined by weight. Input importance depends on the weight; the greater the weight, the more important its value will be. Weight can be negative, and in this case input will be neglected.

The output layer represents the predictable attribute values for the mining model.

12.4.4 Time-Series Algorithm

This algorithm uses a regression technique which is optimized for the forecasting of continuous values. This algorithm does not require an additional column for predicting trends like a decision tree.

Historical information represents the data which is used to create the model, and it is represented on the left of the vertical in Figure 12.6. Predicted information represents the forecasting of the model, and it appears at the right of the vertical.

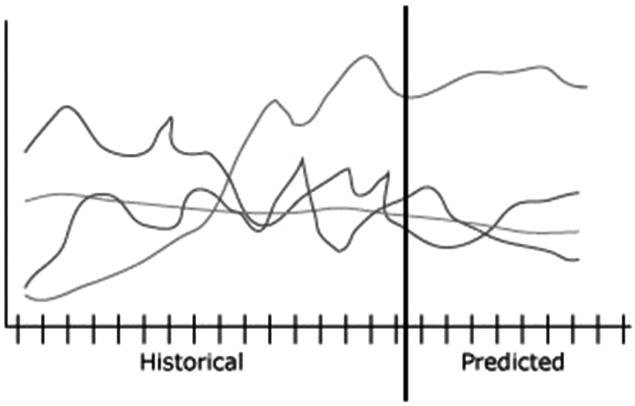


FIGURE 12.6 Time-series trends.

12.4.5 Decision Tree Algorithms

This is an algorithm based on classification and regression, which provides both continuous and discrete predictive modeling. It makes predictions based on the input relation between the columns in the dataset for discrete attributes [29]. For continuous attributes, it will determine where the decision tree splits by using a linear regression technique. If the model includes more than one column, a different decision tree will be formed for each repetitive column. The algorithm trees construct the mining structures by constructing a sequence of divisions in the tree. Splits are viewed as nodes. Every algorithm adds a new node to the column if it predicts a correlation between the input columns. The decision tree algorithm uses a feature selection technique to select attributes which will improve the quality and performance of the mining algorithm. If we use a more predictable attribute, then the model will take a very long time to process, or it will show an out of memory error [30].

12.4.6 Association Algorithm

This algorithm is useful for engines. A suggestion engine recommends items to the consumer or shows the need. The algorithm association [31] is also useful to evaluate the market.

12.5 EXPERIMENTAL ALGORITHM AND RESULTS

In the previous section, we discussed the design specification for performing this research work. The steps involved in the design specification are the basic steps followed in the experimental analysis. In this chapter, we will discuss the various experiments carried out and results obtained from these experiments.

12.5.1 Details of Experiment Carried Out

The experiments carried out to achieve the above-mentioned objectives are described in various stages. From data analysis to integration to reporting, the following experiments were performed.

Stage 1. Environmental Setup:

The above-mentioned hardware and software requirements were fulfilled, and all the software was downloaded. To establish the environment for the experiment, the steps below were followed.

Windows 7 OS was installed on the system.

The visual studio was installed on the machine.

The SQL server 2008 r2 was installed.

We installed SQL server 2008 r2, which included integration, analytical, and reporting services,

Stage 2. Create OLAP Cubes:

The data present in the OLAP system was present in a multi-dimensional structure, and it was created with the help of facts and dimensions [32]. The dimensions had a granularity of viewing data. Therefore, day, month, and year was a TIME dimension hierarchy which specified various aggregation levels. Another dimension which we used was in/outpatient data: Patient age, cancer disease group, sex, diagnosis. We used an OLAP [33] cube creation with the help of facts and dimensions, but it was difficult to find trends and patterns in large OLAP dimensions; therefore, we used data mining techniques.

Step 3. Applying Market Basket Analysis:

We applied algorithms on generated OLAP cubes to extract useful data from the large datasets.

```
D = The whole database
s = Choose a random sample from database D
S = Large_patientsitemsets in s
F = patientsitemsets having >= min_chance_of_cancer
Report if Error return(F)
```

12.5.2 Sampling Algorithm

```
P = Partition Patients_Database(n) n = The number of
Partitions
for (i = 1 ; i ≤ n ; i + +) {
Read from Partition(pi ∈ P)
Li = gen _large_patientsitemsets(pi)
}
```

```
/* Merge Phase */
for (i = 2 ; Lij != Ø , j = 1, 2, . . . , n ; i + +)
{
CGi = Uj=1,2,...,nLij
}
/* 2nd Phase */
for (i = 1 ; i ≤ n ; i + +) {
Read from Partition(pi ∈ P)
forall candidates c ∈ CGgen count(c, pi) }
LG = {c ∈ CG|c.count ≥ min_chance_of_cancer} return(LG);

/* Processing Step */
The processor searches its partitions to find locally wide
supports for patients.

Compute L1 and calculate C2 = Apriori_Gen(L1)

Virtually Prune C2

Initialize the common portion of the rest of the patients

Configure to create a uniform network.
/* Parallel Step: Every processor i runs this in its partition
   Di */

while(some processor has not finished counting the items on
the shared part)
{
  while( processor i has not finished counting the
patientsitemsets in the shared part)
  {
Scan the next interval on Di and count the patients item sets
in the shared part

    Find the locally large patients item sets among the ones
    in the shared part
    Generate new patients from these locally large
    patientsitemsets
    Perform virtual partition pruning and put shared part
    Remove globally small patientsitemsets in the shared part
  }
}

Generate Parallel Patients dataset of different diseases
procedurePatientDieases
 {
          let Patients set L = Ø;
   let People set F = {Ø};
   while (F != Ø)
```

```
                        {
                           let Male set C = Ø;
    forall database tuples t
    {
            forall patientsitemsets f in F
                {
        if (t contains f)  {
                                    let Cf =Male
                                    patientsitemsets that are
                                    extensions of f and
                                    contained in t forall
                                    patientsitemsets cf in Cf
             {
               if(Cf C)   cf .count + +
                 else {
                       cf .count = 0
                       C = C + cf
                       }
              }
                              }
                }
    }
let F = Ø
forall patientsitemsets c in C {
if count(c)/|D| > min_chance_of_cancer
L = L + c
 if c should be used as a People in the next pass F = F + c
}
Find large Patientsitemset or similar data set
```

12.5.3 Results and Discussion

In the previous section, we described the various experiments performed to achieve the objective.

The objective of this chapter consisted of two main parts.

- To determine the cancer site and the morphology patterns between various patients.
- To explore data mining applications and challenges in healthcare.

We chose SQL server 2008 integration, analytical, reporting services. We were able to provide a dependency network for cancer patients. In our research, we explored data mining applications and challenges in healthcare.

Market Basket Analysis

Market basket analysis showed the dependency of the attributes which were strongly correlated. Figure 12.7 shows the market basket analysis of patient data. Here we

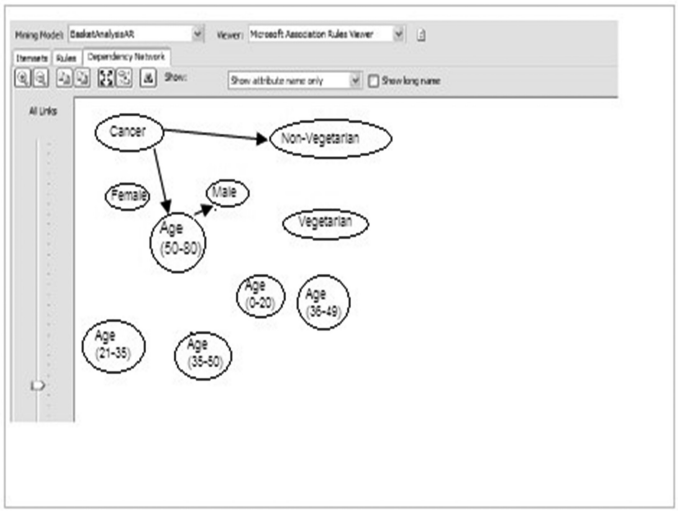


FIGURE 12.7 Market basket analysis on patient data.

found some very interesting patterns in the data. Initially, we defined the attributes for patient analysis, such as age, gender, vegetarian, non-vegetarian. From the above analysis, we observed that patients in the age group of 35–40, male, and non-vegetarian occurred more frequently in the cancer treatment category.

The graph depicted in Figure 12.8 shows the no. of cancer patients in different age groups. The above analytical graph is calculated on eight years of patient data. It

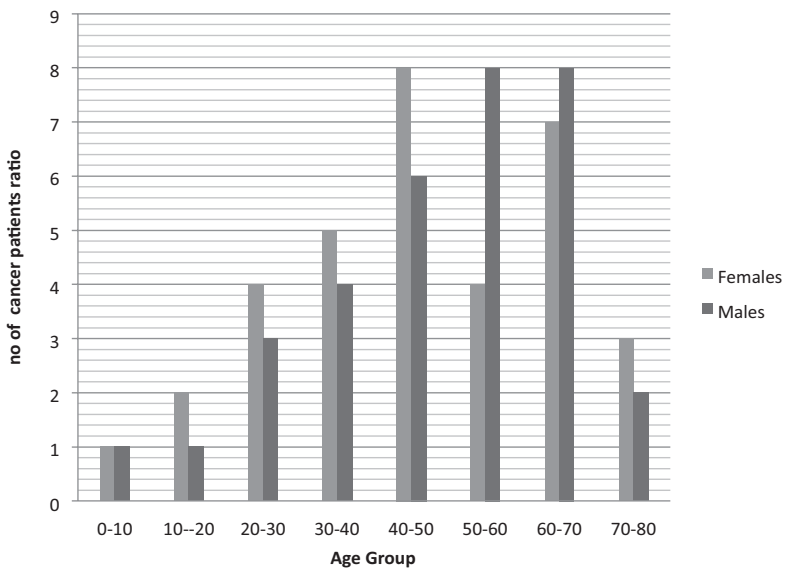
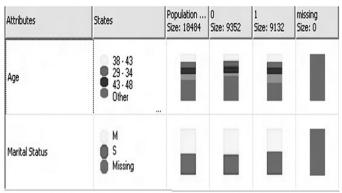


FIGURE 12.8 Graph showing no. of cancer patients in different age groups.

TABLE 12.9
Naïve Bayes Algorithm V/s Market Basket Analysis

Naïve Bayes Algorithm

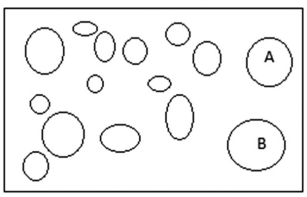
By using a naïve Bayes algorithm we can use predictive modeling. Naïve Bayes algorithm discovers the relation between input and predictable column.



Output screen of naïve Bayes algorithm.

Clustering Algorithm

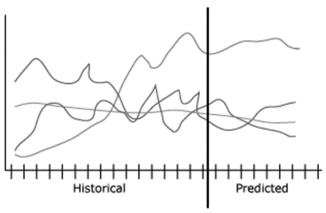
It is a segmentation algorithm used to make clusters which contain similar type of characteristics. Clustering algorithms are used for finding a similar object in data for prediction.



Output screen for a cluster.

Time-Series Algorithm

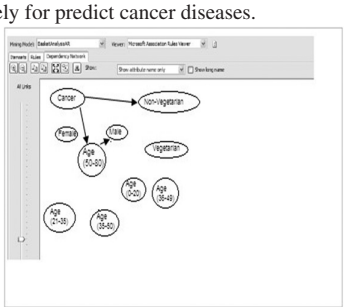
Time-series algorithm uses regression technique which is optimized for the forecasting of continuous value. This algorithm required historical data of patients sets.



Output screen of time series.

Market Basket Analysis

Market basket analysis show dependency on the attributes which are strongest correlated. In my research work, we find very interesting patterns out of different patient's dataset. Firstly we defined the attribute for patient's analysis such as age, gender, veg, and non-veg. From the above analysis, we can see that patients in the age group of 35–40, male, and non-vegetarians are more in the cancer predict. So the result shows people belong this category check their health timely for predict cancer diseases.



Output screen for market basket analysis.

shows very interesting graphs over cancer patients. Most of the cancer patients present between the 40–70 years of age group, and most of them are count of the female patient is more than male patients.

12.5.4 Comparison Study

Table 12.9 shows the comparative study of a naïve Bayes algorithm V/s market basket analysis.

12.6 CONCLUSION AND FUTURE SCOPE

12.6.1 Conclusion

Health issues are arising faster than ever before and scientists are constantly running behind in finding technologies to provide solutions to the new diseases. Data mining is one of the best solutions to treat patients with the help of past experience and knowledge extracted from data collected over years. Data mining has a significant impact in the field of healthcare. Healthcare industries are improving their outputs with the use of various techniques and equipment developed by medical scientists. In this chapter, we proposed to carry out a study of cancer site and morphology patterns among various patients in the context of complex data, such as text, images, sounds and videos. In our approach, we combined OLAP with data mining which resulted in a high level of analysis, and helped us to discover hidden patterns in the data and provide visualization of the information.

12.6.2 Future Scope

This research work provided an association between OLAP and data mining (OLAP mining) for analysis. For future work, we found various issues to address in the future. The first issue was to provide an association between OLAP and predictive analytics using a data mining algorithm. The second was to provide a facility from which doctors could query the data cube on aspects of a business problem and translate this problem into a MDX (multi-dimension expression) query automatically.

REFERENCES

1. Shweta Kharya discussed. 2012. Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, 2(2), pp. 55–66.
2. Patil, D., and Wadhai, V. 2012. Dynamic Data Mining Approach to WMRHM. In *Proceedings of 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Singapore, pp. 1978–1983.
3. Last, M., Carel, R., and Barak, D. 2007. Utilization of Data-Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, Omaha, NE, pp. 169–174. doi: 10.1109/ICDMW.2007.50.

4. Reyes, A. J. O., Garcia, A. O., and Mue, Y. L. 2014. System for Processing and Analysis of Information Using Clustering Technique. *Latin America Transactions, IEEE (Revista IEEE America Latina)* 12(2), 364–371.
5. Goil, S., and Choudhary, A. 1999. A Parallel Scalable Infrastructure for OLAP and Data Mining. In *International Symposium Proceedings on Database Engineering and Applications, IDEAS '99*, Montreal, QC, pp. 178–186.
6. Ahmed, K., Emran, A., Jesmin, T., Mukti, R. F., Zamilur Rahman, Md., and Ahmed, F. 2013. Early Detection of Lung Cancer Risk Using Data Mining. *Asian Pacific Journal of Cancer Prevention: APJCP*, 14(1), 595–598.
7. Krishnaiah, V., Narsimha, G., and Chandra, N. S. 2013. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Science and Information Technologies* 4(1), 39–45.
8. Akay, A., Dragomir, A., and Erlandsson, B.-E. 2013. A Novel Data-Mining Approach Leveraging Social Media to Monitor and Respond to Outcomes of Diabetes Drugs and Treatment. *IEEE Point-of-Care Healthcare Technologies (PHT)* 2013, 264–266.
9. Ariyadasa, S. N., et al. 2013. Knowledge Extraction to Mitigate Child Malnutrition in Developing Countries (Sri Lankan Context). In *4th International Conference on Intelligent Systems, Modelling and Simulation*, IEEE Computer Society, Washington, DC, pp. 321–326.
10. Durairaj, M., et al. 2012. An Empirical Study on Applying Data Mining Techniques for the Analysis and Prediction of Heart Disease. In *International Conference on Information Communication and Embedded Systems (ICICES)*, IEEE, Chennai, pp. 265–270.
11. Bellazzi, R., Sacchi, L., and Concaro, S. 2009. Methods and Tools for Mining Multivariate Temporal Data in Clinical and Biomedical Applications. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5629–5632.
12. Xylogiannopoulos, F. 2012. Developing an Efficient Health Clinical Application: IIOP Distributed Objects Framework. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, pp. 759–764.
13. Santhi, P., and Murali Bhaskaran, V. 2010. Performance of Clustering Algorithms in Healthcare Database. *International Journal for Advances in Computer Science* 2(1), 26–31.
14. Ha, S. H., and Joo, S. H. 2010. A Hybrid Data Mining Method for the Medical Classification of Chest Pain. *World Academy of Science, Engineering and Technology, Open Science Index 37, International Journal of Computer and Information Engineering* 4(1), 99–104.
15. Yan, H., Zheng, J., Jiang, Y., Peng, C., and Li, Q.. 2003. Development of a Decision Support System for Heart Disease Diagnosis Using Multilayer Perceptron. In *Proceedings of the 2003 International Symposium on Circuits and Systems (ISCAS) '03.* 5, pp. V–V.
16. Andreeva, P. 2006. Data Modelling and Specific Rule Generation via Data Mining Techniques. In *International Conference on Computer Systems and Technologies—CompSysTech*, pp. IIIA.17-1–IIIA.17-6.
17. Hara, A., and Ichimura, T. 2008. Data Mining by Soft Computing Methods for the Coronary Heart Disease Database. In *IEEE Fourth International Workshop on Computational Intelligence & Application*, *SMC Hiroshima Chapter*, Hiroshima University, Japan, 10-11.
18. Sitar-Taut, V. A. 2009. Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation. *Journal of Applied Computer Science & Mathematics* 3(1), 29–32.
19. Chang, C. L., and Chen, C. H. 2009. Applying Decision Tree and Neural Network to Increase Quality of Dermatologic Diagnosis. *Expert Systems with Applications, Elsevier* 36, 4035–4041.

20. Rajkumar, A. and Reena, G. S. 2010. Diagnosis of Heart Disease Using Data mining Algorithm. *Global Journal of Computer Science and Technology* 10(10), pp 38-43.
21. Srinivas, K., Rani, B. K., and Govrdhan, A. 2010. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering (IJCSE)* 02(02), 250–255.
22. Anbarasi, M., Anupriya, E., and Iyengar, N. 2010. Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm. *International Journal of Engineering, Science and Technology* 2(10), 5370–5376.
23. Kangwanariyakul, Y., Nantasenamat, C., Tantimongcolwat, T., and Naenna, T. 2010. Data Mining of Magnetocardiograms for Prediction of Ischemic Heart Disease. *EXCLI Journal*, 9, 82–95, 2010.
24. Osareh, A., and Shadgar, B. 2010. Machine Learning Techniques to Diagnose Breast Cancer. In *5th International Symposium on Health Informatics and Bioinformatics*, pp. 114–120.
25. Abdi, M. J., and Giveki, D. 2013. Automatic Detection of Erythemato-Squamous Diseases Using PSO–SVM Based on Association Rules. *Engineering Applications of Artificial Intelligence* 26, 603–608.
26. Kumar, A., Goyal, D., and Dadheech, P. 2018. A Novel Framework for Per-formance Optimization of Routing Protocol in VANET Network. *Journal of Advanced Research in Dynamical & Control Systems* 10, 2110–2121, ISSN: 1943-023X.
27. Dadheech, P., Goyal, D., Srivastava, S., and Kumar, A. 2018. A Scalable Data Processing Using Hadoop & MapReduce for Big Data. *Journal of Advanced Re-search in Dynamical & Control Systems* 10, 2099–2109, ISSN: 1943-023X.
28. Dadheech, P., Goyal, D., Srivastava, S., and Choudhary, C. M.. 2018. An Efficient Approach for Big Data Processing Using Spatial Boolean Queries. *Journal of Statistics and Management Systems (JSMS)* 21(4), 583–591.
29. Kumar, A. and Sinha, M. 2014. Overview on Vehicular Ad Hoc Network and Its Security Issues. In *International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 792–797. doi: 10.1109/IndiaCom.2014.68280712.
30. Dadheech, P., Kumar, A., Choudhary, C., Beniwal, M. K., Dogiwal, S. R., and Agarwal, B. 2019. An Enhanced 4-Way Technique Using Cookies for Robust Authentication Process in Wireless Network. *Journal of Statistics and Management Systems* 22(4), 773–782. doi: 10.1080/09720510.2019.1609557.
31. Kumar, A., Dadheech, P., Singh, V., Raja, L., and Poonia, R. C. 2019. An Enhanced Quantum Key Distribution Protocol for Security Authentication. *Journal of Discrete Mathematical Sciences and Cryptography* 22(4), 499–507. doi: 10.1080/09720529.2019.1637154.
32. Kumar, A., Dadheech, P., Singh, V., Poonia, R. C., and Raja, L. 2019. An Improved Quantum Key Distribution Protocol for Verification. *Journal of Discrete Mathematical Sciences and Cryptography* 22(4), 491–498. doi: 10.1080/09720529.2019.1637153.
33. Kumar, A., and Sinha, M. 2019. Design and Analysis of an Improved AODV Protocol for Black Hole and Flooding Attack in Vehicular Ad-Hoc Network (VANET). *Journal of Discrete Mathematical Sciences and Cryptography* 22(4), 453–463. doi: 10.1080/09720529.2019.16371512.