Diabetic data analysis in healthcare using Hadoop architecture over big data

P. Sampath*

Department of CSE,
Sasurie College of Engineering,
Vijayamangalam, Perundurai, Tamil Nadu, India
Email: samkala@gmail.com
*Corresponding author

S. Tamilselvi

Department of Biotechnology,
Bannari Amman Institute of Technology,
Sathyamangalam, Tamil Nadu, India
Email: tamilselvis@bitsathy.ac.in

N.M. Saravana Kumar

Department of IT,
Vivekanandha College of Engineering for Women,
Tiruchengode, Tamil Nadu, India
Email: saravanakumaar2008@gmail.com

S. Lavanya and T. Eswari

Department of IT,
Sri Krishna College of Engineering and Technology,
Coimbatore, Tamil Nadu, India
Email: lavanyamecsbit@gmail.com
Email: eswarit@skcet.ac.in

Abstract: Owing to the increasing developments in this digitised era, it is necessary to move from paper health records to digital by handling the large volume of healthcare data for analysis, and using them for efficient treatment will be a crucial issue. Diabetes Mellitus (DM) is one of the Non-Communicable Diseases (NCD). It is a major health hazard in developing countries and associated with long-term complications and numerous health disorders. The main idea of this project is to integrate massive unstructured diabetic data from various sources which need to be normalised into a proper scale to get optimised solution for medicinal field using Hadoop. Although the traditional database management systems can handle data effectively, processing the high volume of unstructured data at a reasonable time becomes very challenging. This paper uses the predictive analysis algorithm in the Hadoop/MapReduce environment to predict the DM complexities and the type of treatment to be adopted. Based on the analysis, this system provides an efficient way to cure patients with better outcomes.

Copyright © 2017 Inderscience Enterprises Ltd.

Keywords: predictive analysis; Hadoop MapReduce; healthcare; diabetes mellitus; big data analytics.

Reference to this paper should be made as follows: Sampath, P., Tamilselvi, S., Saravana Kumar, N.M., Lavanya, S. and Eswari, T. (2017) 'Diabetic data analysis in healthcare using Hadoop architecture over big data', *Int. J. Biomedical Engineering and Technology*, Vol. 23, Nos. 2/3/4, pp.137–147.

Biographical notes: P. Sampath completed his undergraduate degree in Electronics and Communication Engineering at Bharathiyar University, Coimbatore in 1991. He completed his postgraduate in Computer Science Engineering at Bharathiyar University, Coimbatore in 2001 and completed his PhD in Information and Communication Engineering at Anna University, Chennai in 2015. Currently, he is Professor and Head of the Department of Computer Science and Engineering in Sasurie College of Engineering College at Tamilnadu. He is having more than 22 years of experience in teaching and industry.

S. Tamilselvi is working as an Associate Professor in the Department of Bio Technology in Bannari Amman Institute of Technology, Erode. She completed her PhD in science major at Anna University, Chennai. She completed her MPhil from Bharathiyar University. She completed her undergraduate and postgraduate from Avinashilingam, University, Coimbatore. She has 11years of academic experience and published more than 10 papers in various international journals.

N.M. Saravana Kumar is currently working as a Professor and Head of the Department in the Department of IT, Vivekanandha College of Engineering for Women, Nammakkal, Tamil Nadu, India. He completed his PhD in Network Security at Anna University, Chennai. He has published 20 papers in various reputed journals. He has 14 years of academic experience, and his research interests include parallel computing, distributed systems and network security.

S. Lavanya is currently working as an Assistant Professor in the Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India. She received her master degree in Computer Science Engineering. She has published more than ten papers in various international journals. Her areas of interest include network security, wireless networks, cloud computing.

T. Eswari is currently working as an Assistant Professor in the Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India. She received her master degree in Computer Science Engineering. She has published three papers in various international journals and presented seven papers in various national and international conferences across India. Her areas of interest include cloud computing and wireless sensor networks.

1 Introduction

In healthcare, data mining is becoming very popular because its applications can significantly benefit all parties involved in the healthcare industry. For example, data

mining can help healthcare insurers detect fraud and abuse, healthcare organisations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better, affordable healthcare services (Koh and Tan, 2005).

In recent years, healthcare industry has generated large amounts of data. The value-based treatment in hospitals and digitisation of world call for computerised data rather than hard copies. The healthcare data include Electronic Health Reports (EHR) of patients data, clinical reports, doctor's prescription, diagnostic reports, medical images, pharmacy information, health-insurance-related data, data from social media and medicinal journals (Saravana Kumar et al., 2015; Kushniruk, 2008). All these information collectively forms big data in healthcare.

Employing the analysis of big data will produce the predicted results for understanding the trends to improve healthcare and life expectancy and proper treatment at early stages at low cost. The analytics associated with big data are described by four characteristics: volume, velocity, variety and veracity (Raghupathi and Raghupathi, 2014). The accumulation of health-related data continuously results in an incredible volume of data. Velocity is accessing those data in real time at a rapid speed. Variety includes diabetic glucose measurements, blood pressure readings and various EHRs. Veracity assumes the simultaneous scaling up in performance of the architectures and platforms, algorithms and tools to match the need of big data (Raghupathi and Raghupathi, 2014). The healthcare industry is moving from reporting facts to discovery of insights, towards data-driven healthcare organisations. Big data holds great potential to change the whole healthcare value chain from drug analysis to patients caring quality.

The probability of a 30- to 70-year-old Indian dying from the four main non-communicable diseases (NCD) – diabetes, cancer, stroke and respiratory diseases – is 26% at present, according to the World Health Organisation (Saravana Kumar et al., 2015). According to the Global Status Report, NCDs would claim nearly 52 million lives globally by the year 2030. According to WHO, nearly 8.5 million people died of NCDs in the South-East Asia Region in 2012. In India, NCDs are estimated to have accounted for 60% of all deaths in 2014, while 26% between the ages of 30 and 70 years had a probability of succumbing to one of the four diseases.

Diabetes Mellitus (DM), one of the NCDs, is a major health hazard in developing countries such as India. The acute nature of DM is associated with long-term complications and numerous health disorders (Saravana Kumar et al., 2015). There are three main types of this disease. Type 1 DM results from the body's failure to produce insulin, and presently requires the person to inject insulin. This form is referred as Insulin-Dependent Diabetes Mellitus (IDDM). Type 2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as Non-Insulin-Dependent Diabetes Mellitus (NIDDM).

The third main form, gestational diabetes, occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2 DM. It was estimated that 61.3 million people aged 20–79 years live with diabetes as of 2011 in India. This number is expected to increase to 101.2 million by 2030.

Due to the growing unstructured nature of diabetic data from health industry or all other sources, it is necessary to structure and emphasise its size into nominal value with possible solution (Saravana Kumar et al., 2015). With the help of technological

developments, it is necessary to combine robust diabetic data sharing and electronic communication systems which facilitate better access to health services at all the levels of patients. Therefore, all patient information needs to be in one repository.

Deploying Health Information Exchange (HIE) can extract clinical information from several disparate repositories and integrate those data within a single patient health record that all care providers can access securely. Predictive analysis is a method that incorporates a variety of techniques from data mining, statistics, and game theory that uses the current and past data with statistical or other analytical models and methods to determine or predict certain future events (Mishra and Silakari, 2012). Significant predictions or decisions can be made by employing big data analytics in healthcare field. In this paper, we use the predictive analysis algorithm in Hadoop/MapReduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability.

2 Reviewed schemes

A literature review reveals many results on diabetes carried out by different methods and materials of diabetes problem in India. Many people have developed various prediction models using data mining to predict diabetes (Saravana Kumar et al., 2015). Combination of classification, regression, genetic and neural network handles the missing and outlier values in the diabetic data set, and also replaces the missing values with the domain of the corresponding attribute (Bhat et al., 2009). The classical neural network model is used for prediction on the pre-processed data set.

In predictive analysis of diabetic treatment using regression-based data mining techniques to diabetes data, they discover patterns using SVM algorithm that identify the best mode of treatment for diabetes across different ages (Aljumah et al., 2013). They concluded that drug treatment for patients in the young age group can be delayed, whereas patients in the old age group should be prescribed drug treatment immediately. Prediction and classification of various type of diabetes using C4.5 classification algorithm was carried out in Pima Indians Diabetes Database (Rajesh and Sangeetha, 2012). A detailed analysis of the Pima diabetic data set was carried out efficiently using Hive and R. In this analysis, deriving some interesting facts provides way to develop prediction models (Sadhana, 2014).

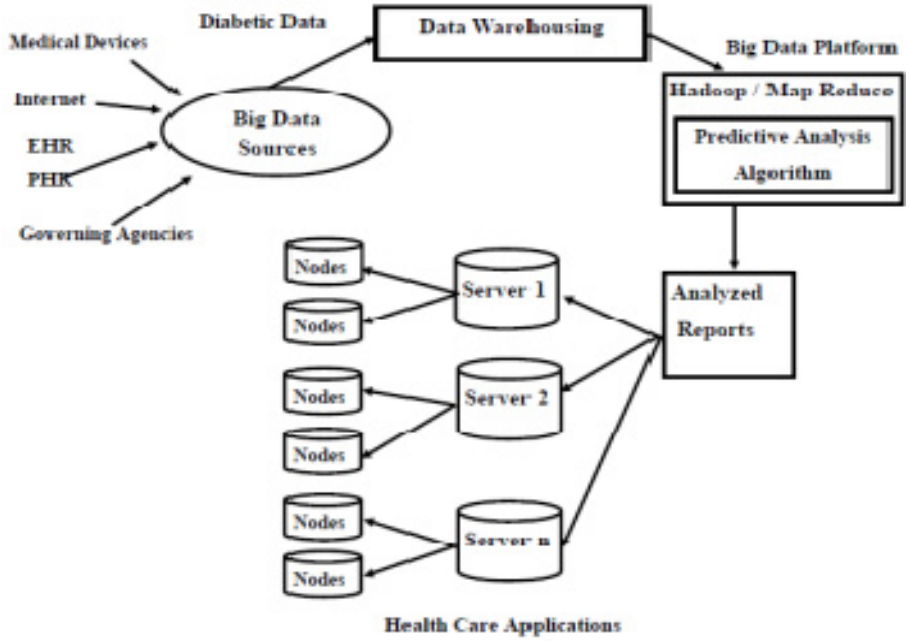
The soft computing-based prediction model was developed for finding the risks accumulated by the diabetic patients. They have experimented with real-time clinical data using genetic algorithm (Sabibullah et al., 2013). The obtained results pertain to the level of risk prone to either heart attack or stroke. The novel pre-processing phase is missing value imputation for both numerical and categorical data. A hybrid combination of Classification and Regression Trees (CART) and genetic algorithms to impute missing continuous values and Self-Organising Feature Maps (SOFM) to impute categorical values was improved in Bhat et al. (2011).

All the above researches have been successful in analysing the diabetic data set and developing good prediction models. In this paper, we use the predictive analysis technique in Hadoop/MapReduce environment to predict and classify the type of diabetes. This system provides efficient way to care and cure patients at low cost with better outcomes like affordability and availability.

3 Predictive analysis system architecture

The proposed system integrates all internal, external and multi-format data collected from various sources. The architecture of predictive analysis system includes various phases like data collection, data warehousing, predictive analysis and processing analysed reports. Figure 1 shows the complete architecture of the proposed method.

Figure 1 Architecture of the predictive analysis system



3.1 Data collection

The raw diabetic big data or data set are given as input to the system (Saravana Kumar et al., 2015). The unstructured voluminous input data can be obtained from various Electronic Health Record (EHR)/Patient Health Record (PHR), clinical systems and external sources (government sources, laboratories, pharmacies, insurance companies, etc.), in various formats (flat files, .csv, tables, ASCII/text, etc.) and residing at various locations (Raghupathi and Raghupathi, 2014). Those raw data can be transformed into useful facts by using a data processing tool.

3.2 Data warehousing

In this phase, massive unstructured data are warehoused into single unit in which data from various sources are cleansed, accumulated and made ready for further processing (Saravana Kumar et al., 2015). Integration of various EHRs can help in identifying the patterns for diabetes prediction system.

3.3 Predictive analysis

Predictive analysis can help healthcare providers accurately expect and respond to the patient needs (Saravana Kumar et al., 2015). It provides the ability to make financial and clinical decisions based on predictions made by the system. This system uses the predictive analysis algorithm in Hadoop/MapReduce environment to predict and classify the type of DM, complications associated with it and the type of treatment to be provided.

3.3.1 Hadoop

Hadoop is the open-source distributed data processing platform from Apache. Hadoop can serve the twin roles of data organiser and analytics tool (Raghupathi and Raghupathi, 2014). Hadoop has the potential to process extremely large amounts of health data mainly by allocating partitioned data sets to numerous servers like clusters, each of which solves different parts of the larger problem and then integrates them for the final result (Saravana Kumar et al., 2015). Hadoop uses two main components to do its job: MapReduce and Hadoop Distributed File System (HDFS).

- *MapReduce*: Hadoop's implementation of MapReduce is based on programming models to process large data or data sets by dividing them into small blocks of tasks. MapReduce uses distributed algorithms, on a group of computers in a cluster, to process large data sets. It consists of two functions:
    - The Map ( ) function which resides on the master node and then divides the input data or task into smaller subtasks, which then distributes to worker nodes that process the smaller tasks and pass the answers back to the master node. The subtasks are run in parallel on multiple computers.
    - The Reduce ( ) function collects the results of all the subtasks and combines them to produce an aggregated final result, which in turn is the answer to the original big query.
- *HDFS*: HDFS replicates the data blocks that reside on other computers in the data centre (to ensure reliability) and manages the transfer of data to the various parts of the distributed system.

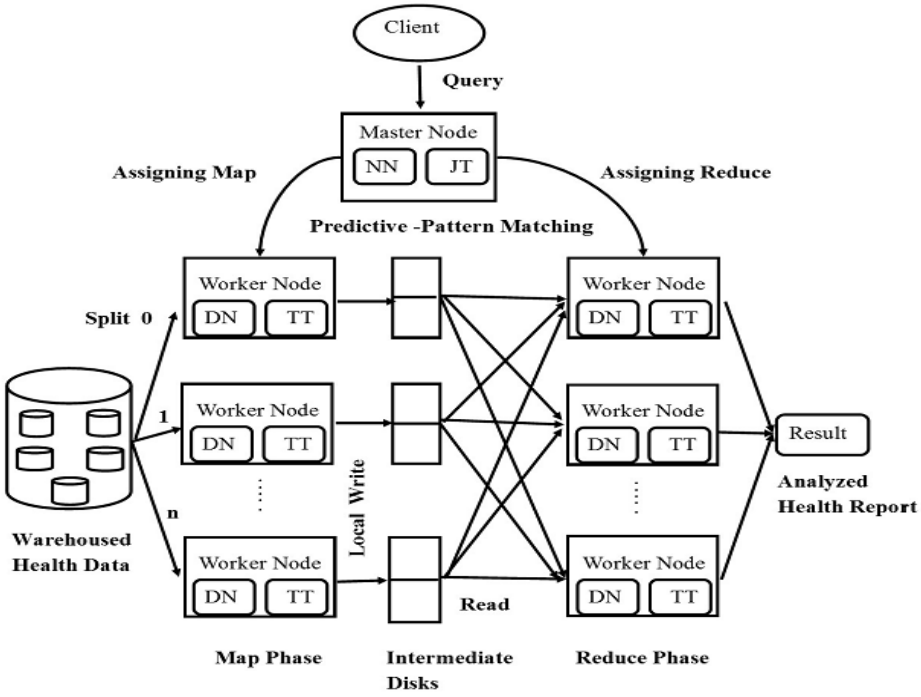
3.3.2 Pattern discovery

For diabetic treatment, it is necessary to test patterns like plasma glucose concentration, serum insulin, diastolic blood pressure, diabetes pedigree, Body Mass Index (BMI), age and number of times a patient has been pregnant (Saravana Kumar et al., 2015). The pattern discovery of predictive analysis must include the following (http://www.intel.com/content/www/us/en/healthcare-it/solutions/documents/bigger-data-better-healthcare-idc-insights-white-paper.html):

- *Association rule mining*: association between diabetic type and pages viewed (e.g. laboratory results).
- *Clustering*: clustering of similar patterns of usage, etc.

- *Classification*: classification of health risk value by the level of patient health condition.
- Usage of statistics.
- Application of pre-defined deductive rules across data.

Figure 2 Predictive – pattern matching system



3.3.3 Predictive pattern matching

Whenever the warehoused data set was sent to Hadoop system, immediately the map reduce task is performed. In the mapping phase, the Master Node splits large data into smaller tasks for numerous Worker Nodes. Figure 2 deploys the exact operation of predictive pattern matching system. The Master Node consists of Name Node (NN) and Job Tracker (JT), which always employs the map and reduce task (Saravana Kumar et al., 2015). The Worker Node or Slave Node receives the order from the Master Node and processes the pattern matching task for diabetes data with the help of Data Node (DN), Same Machine and Task Tracker (TT).

Predictive matching is the process of comparing the analysed threshold value with the obtained value. If the pattern matching process was completed by all Worker Nodes based on the requirement, it was stored in intermediate disks. This process is known as local write. If the reduce task was initiated by Master Node, all other allocated Worker

Nodes will read the processed data from intermediate disks. Based on the query received from Client through Master Node, the reduce task will be performed in Worker Node. The results obtained from the reduce phase will be distributed in various servers.

3.4 Processing analysed reports

After the analysis of large diabetic data went through Hadoop, the final results are distributed over various servers and replicated through several nodes depending on the geographical area (Saravana Kumar et al., 2015). Employing proper electronic communication technology to exchange the information of individual patients among healthcare centres will lead to proper treatment at the right time in remote locations at low cost.

4 Experimental results

4.1 Data set analysis

In this paper, the data set we used for diabetic data analysis is Pima Indians Diabetes Database of National Institute of Diabetes, Digestive and Kidney Diseases. This diabetes database, donated by Vincent Sigillito, is a collection of medical diagnostic reports of 768 examples from a population living near Phoenix, Arizona, USA. The data set samples consist of values with eight attributes and one of the two possible outcomes, namely whether the patient is tested positive for diabetes (indicated by 1) or not (indicated by 0). This data set is analysed using the predictive analysis approach using Hadoop MapReduce environment.

Once the file is loaded to Hadoop to perform pattern discovery for the pattern matching process in the data set. Previously the necessary threshold values of diabetic diagnosis attributes were collected from various hospitals. Then the data set was partitioned based on the attributes. There are eight attributes in the data set and for each data set, the number of distinct values was found out and the time taken to perform this analysis was recorded. This experimentation was performed in 10 runs.

In each run, a set of 40 different queries (i.e. jobs) have been executed with different runs. In each run, the execution time and query throughput have been taken and produced the number of requests handled by each agent. Processing of each query is done based on the time it arrives. The following graphs and parameters specify the effectiveness of the idea.

During the analysis, we found that out of 768 records in the data set, 500 records belonged to class 0 (non-diabetic) and 268 belonged to class 1 (diabetic). An analysis of the predicted report of diabetic data set is given in Table 1.

Table 1 Predicted report of diabetic data set

Records in data set	*Obtained result*
Actual records	768
Non-diabetic	500
Diabetic	268

Computation of accuracy rates:

1 Rate of classifying actual diabetic cases correctly:

Sensitivity = 268/768 = 34.99%.

2 Rate of classifying actual non-diabetic cases correctly:

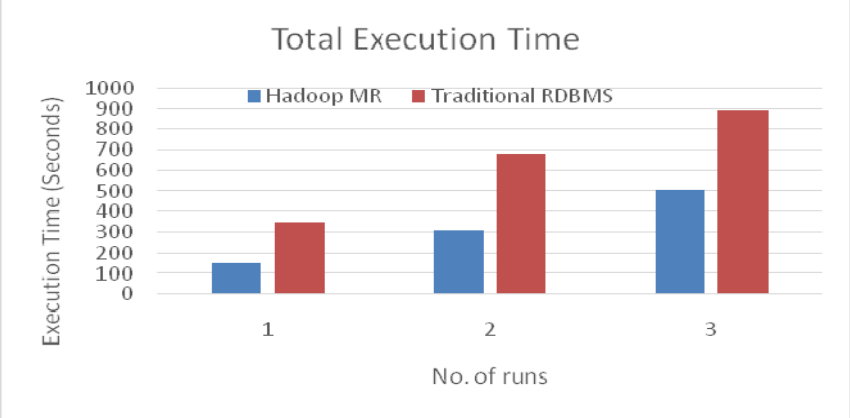
Specificity = 500/768 = 65.19%.

The accuracy rates for predicting diabetic cases are lower but can still be estimated to be sufficient. In particular, for the 268 diabetic cases in the data set, the attributes in the data set was not sufficient for further classification of diabetic types.

4.2 Total execution time

Execution time of queries depends upon the size of the input of data of the data set; for small size of data set, the execution time of traditional DBMS is lower than the Hadoop. The Hadoop analysis set-up is useful for large data sets. Figure 3 shows the total execution time during each run between traditional DBMS and Hadoop MapReduce. For Pima Indians Diabetes data set, during each run, the execution time increases significantly for traditional DBMS rather than Hadoop MapReduce.

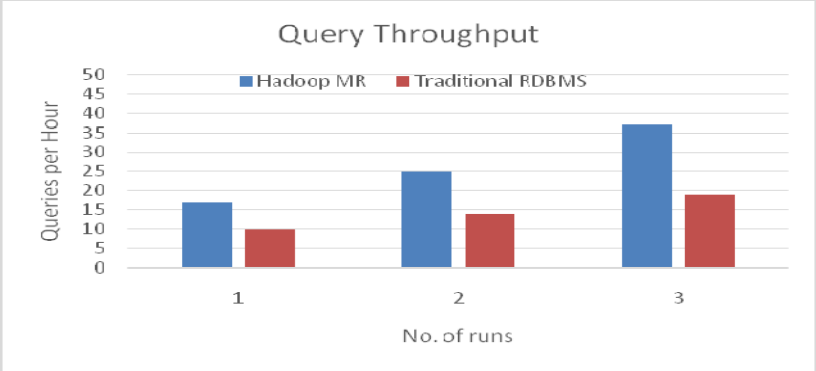
Figure 3 Analysis on total execution time



4.3 Query throughput

The performance of the system is mainly calculated with respect to a specific workload using query throughput (in queries per hour). In each run, a set of 40 different queries (i.e. jobs) have been executed with different runs. Figure 4 shows the query throughput during each run between traditional DBMS and Hadoop MapReduce. For diabetes data set, during each run, the query throughput increases significantly for Hadoop MapReduce rather than traditional DBMS.

Figure 4 Analysis on query throughput



4.4 Benefits of this predictive analysis system

The diabetes may be associated with severe diseases such as heart attacks, strokes, eye diseases and kidney diseases. Analysing the risk value by the level of patient health condition using the above results can be used by the physicians at remote locations to serve the people. This helps in detecting diseases at earlier stages when they can be treated more easily and effectively, which reduces the spending of the government each year for a particular disease (Saravana Kumar et al., 2015). In developing countries such as India, it is mandatory to manage specific individual and population health and detect healthcare fraud more quickly. The middle-income families can be treated with the high availability of medical facility at minimum cost. This system leads to the improved focus on patient-centric treatment, thereby reducing and saving the next generation from DM.

5 Conclusion

The novelty and significance of this work is that by introducing the idea of efficient information utilisation to develop a distributed diabetic data analytics system, this work advances the state of the art in the healthcare arena. Data with more complexities keep evolving in healthcare, thus leading to more opportunities for big data analytics. Big data analytics in Hadoop's implementation provide an efficient way for achieving healthcare service to all population with better outcomes. Diabetes is one of the NCDs and is a major health hazard in developing countries. The proposed system is capable of handling textual information of diabetic data from various sources and it can be extended in such a way to handle the medical image files which further enhances the medical services with better outcomes. The large volume of data can be distributed across multiple nodes in order to process it faster than the existing system. The usage of Hadoop in this system improves the data retrieval speed six times more than the traditional Medicare system that uses the DBMS backbone. The processing of queries performed by predictive analysis algorithm facilitates throughput, minimises healthcare errors and eradicates patient readmission.

References

Aljumah, A.A., Ahamad, M.G. and Siddiqui, M.K. (2013) 'Application of data mining: diabetes health care in young and old patients', *Journal of King Saud University – Computer and Information Sciences*, Vol. 25, No. 2, pp.127–136.

Bhat, V.H., Rao, P.G. and Shenoy, P.D. (2009) 'An efficient prediction model for diabetic database using soft computing techniques', *Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, 15–18 December, Delhi, India, pp.328–335.

Bhat, V.H., Rao, P.G., Krishna, S. and Shenoy, P.D. (2011) 'An efficient framework for prediction in healthcare data using soft computing techniques', *Proceedings of the 1st International Conference on Advances in Computing and Communications*, 22–24 July, Kochi, India, pp.522–532.

Koh, H.C. and Tan, G. (2005) 'Data mining applications in healthcare', *Journal of Healthcare Information Management*, Vol. 19, No. 2, pp.64–72.

Kushniruk, A.W. (2008) *Predictive Analytics and Forecasting in Health Care: Integrating Analytics with Electronic Health Records*, SAS Institute Inc., Cary, NC.

Mishra, N. and Silakari, S. (2012) 'Predictive analytics: a survey, trends, applications, opportunities & challenges', *International Journal of Computer Science and Information Technologies*, Vol. 3, No. 3, pp.4434–4438.

Raghupathi, W. and Raghupathi, V. (2014) 'Big data analytics in healthcare: promise and potential', *Health Information Science and Systems*, Vol. 2, No. 3, pp.2–10.

Rajesh, K. and Sangeetha, V. (2012) 'Application of data mining methods and techniques for diabetes diagnosis', *International Journal of Engineering and Innovative Technology*, Vol. 2, No. 3, pp.224–229.

Sabibullah, M., Shanmugasundaram, V. and Raja Priya, K. (2013) 'Diabetes patient's risk through soft computing model', *International Journal of Emerging Trends & Technology in Computer Science*, Vol. 2, No. 6, pp.60–65.

Sadhana, S.S. (2014) 'Analysis of diabetic data set using Hive and R', *International Journal of Emerging Technology and Advanced Engineering*, Vol. 4, No. 7, pp.626–629.

Saravana Kumar, N.M., Eswari, T., Sampath, P. and Lavanya, S. (2015) 'Predictive methodology for diabetic data analysis in big data', *Elsevier Procedia Computer Science*, Vol. 50, pp.203–208.