ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/359461868

Symptoms-Based Disease Prediction Using Big data Analytics

Chapter · January 2022

DOI: 10.1007/978-981-16-8987-1_36

CITATIONS
3

READS
2,702

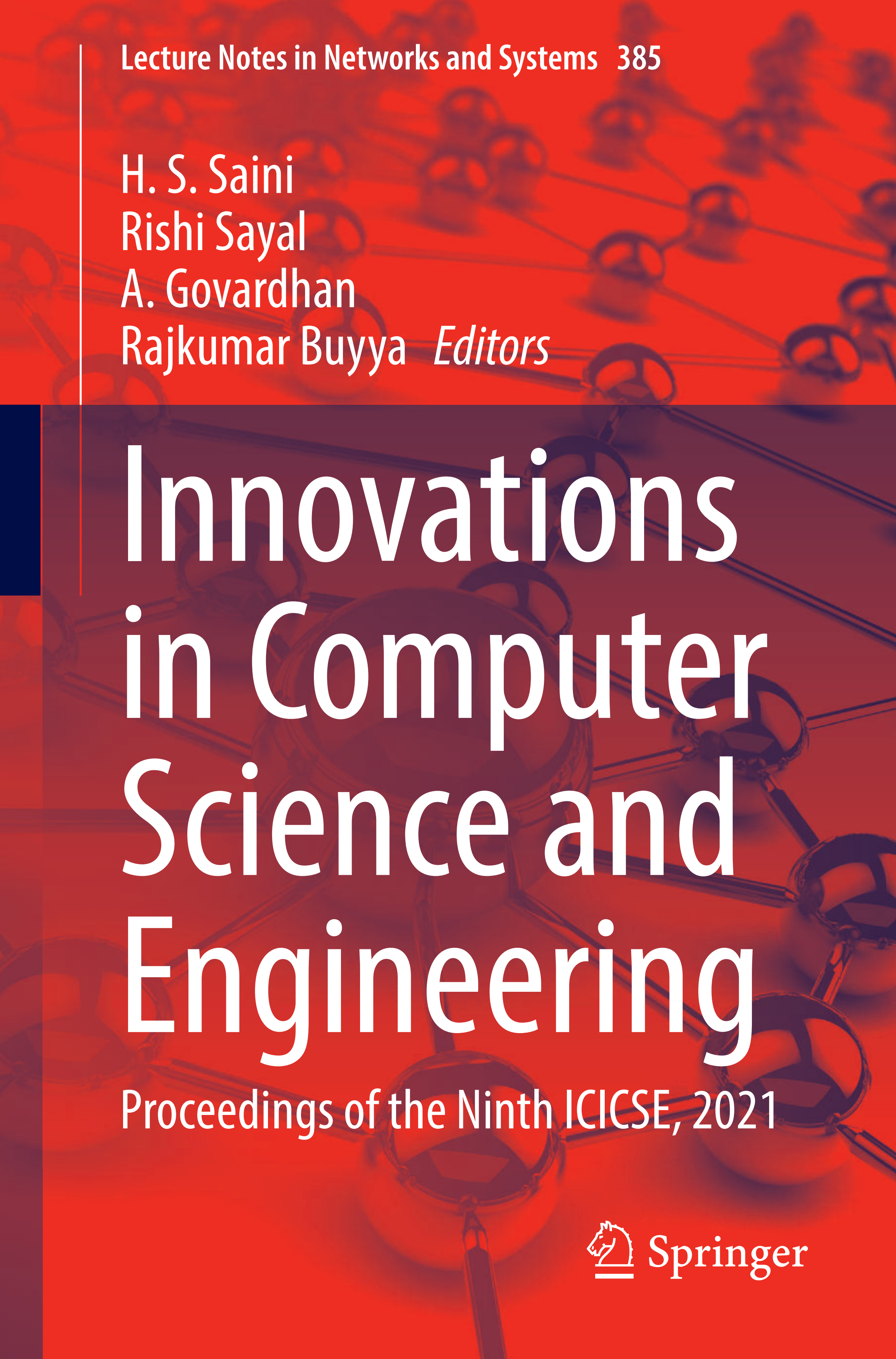
3 authors, including:

Smritilekha Das
Koneru Lakshmaiah Education Foundation
19 PUBLICATIONS **32** CITATIONS

SEE PROFILE

All content following this page was uploaded by Smritilekha Das on 07 April 2022.

The user has requested enhancement of the downloaded file.



Lecture Notes in Networks and Systems 385

H. S. Saini
Rishi Sayal
A. Govardhan
Rajkumar Buyya *Editors*

Innovations in Computer Science and Engineering

Proceedings of the Ninth ICICSE, 2021

Springer

Lecture Notes in Networks and Systems

Volume 385

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Turkey

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Symptoms-Based Disease Prediction Using Big data Analytics

P. Kanchanamala, Smritilekha Das, and G. Neelima

Abstract The data is being collected from various sources in the world which belongs to healthcare sector and is processed called healthcare data management. The huge amount of healthcare data must be processed by efficient tools and methods to create value. In technical market, many BI are available to handle structured data only. But the unstructured data is also being generated which can be used to give valuable insights to improve the quality in healthcare. For understanding the patient needs, there is a need to collect structured and unstructured data from various stakeholders. Then, the analysts get the whole idea about the patient's needs based on symptoms and able to give precision driven care and treatment. The final treatment depends on the patient's present condition and earlier treatment which increases the perfectness in the treatment. This paper the consideration of 400 symptoms and 147 diseases. It analysis, the performance of the machine learning algorithms including Decision tree, Random forest, Naïve Bayes and the proposed algorithm.

Keywords Disease · Symptom · Decision tree · Random forest · Naïve Bayes

1 Introduction

Big data shows its significance in every field in the world including healthcare industry. It changes the way to handle the patients and doctors with care. From more number of sample data, can expect more accurate insights for healthcare industry. Like many industries, healthcare industry is a framework which contains

P. Kanchanamala (✉)
IT Department, GMRIT, Rajam, Andhra Pradesh, India
e-mail: kanchanamala.p@gmrit.edu.in

S. Das
Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur, Andhra Pradesh, India

G. Neelima
Department of CSE, Vignan Institute of Information Technology (Autonomous), Besides VSEZ, Duvvada, Gajuwaka, Visakhaptnam, Andhra Pradesh, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

H. S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Lecture Notes in Networks and Systems 385,
https://doi.org/10.1007/978-981-16-8987-1_36

heterogeneous sectors are complex to handle with high accuracy, where the patients demanding better care with less price. Day by day, new technologies are being included to the healthcare industry, where the big data analytics plays a vital role for giving effective business insights to the hospitals as well as patients.

In the technical world, data analysis plays an important role in every field in the world where the data volume is so limited. But today, the world is in big data era. The existing statistics says that the data analytics is very important in near future for healthcare industry and it becoming very crucial in clinical, operational and financial sectors.

The collected data can potentially be used by the Govt. and public organizations create or improve policies, procedures and trainings. Overall, the project has the potential to heighten awareness for the need to give best treatment in any healthcare environment.

Most of the patients are illiterates, and those are not familiar precision treatment. So majority of people approaching private healthcare centers which are not able to store the details of patients and their diseases. So there is a need to organize health camps which educates and sensitize the community. This framework explains about diagnosis and various types of health hazards.

The objectives of the proposed algorithm are as follows:

- To map high-risk areas for disease prevention
- To devise the framework for sharing Electronic Health Records (EHRs) via secure information systems
- To devise dynamic descriptive decision tool for real-time alerting to design security enhanced features.
- Telemedicine—It is a process to provide the customized treatment for each patient for avoiding re-admission in hospital again and again.

2 Literature Survey

Bates et al. [1, 2], big data analytics can help early disease detection, deviation from healthy state and detection of fraud. It also helps in getting accurate predictions, cost-reduction in healthcare maintenance, and it provides precision good health. Dencelin and Ramkumar [3] proposed a framework for analyzing big data with the help of Apache Spark.

Fang et al. [4] proposed a framework titled "Health informatics processing pipeline framework" which consists of data capturing, storing, analyzing, searching and decision support. It offers dynamic services to the patients through mobile devices and sensor networks. Legaz Garca et al. [5] proposed OWL based framework which gathers patient data (EHR) and utilized for data exploration.

The framework [6] has been proposed for healthcare system, has four layers. The advantages of this framework are data optimization and data security. It is based on distributed model and enhances the performance of the system by data and storage optimization.

The framework which contains layers has been proposed for healthcare system by Raghupathi and Raghupathi [7]. The data source layer handles internal and external data sources for healthcare system. The transformation layer for transformation and loading the data. The analytics layer for querying, reporting and processing. Theoretically, these concepts are good enough.

Sakr and Elgammal [8] proposed a method that integrates sensors, cloud, IoT. It is able to handle patient profile analytics, population management, etc. Pramanik et al. (2017) proposed a layered framework on healthcare system. This framework yields useful smart system services.

The data processing concept Sunil Kumar et al. [9] has been explained. The healthcare data is being generated and coming from various sources in the form of EHRs, genome database, text and imagery unstructured data, clinical reports, sources belongs to Govt. sector, lab reports from medical centers and pharmacies and health insurance companies. This data can be handled by HADOOP framework.

After data collection, predictions can be done using machine learning algorithms. For healthcare industry data, two types of prediction algorithms supervised and unsupervised are required.

3 Methodology

Data Collection

The data can be collected from various repositories and store the entire data in Hadoop Distributed File System (HDFS). The data can be collected through surveys and questionnaires, focus groups, interviews, and observations and progress tracking.

Sources of Data

Data can be collected from different sources like hospitals, medical practitioners, patient health history, surveys, medical bills, etc.

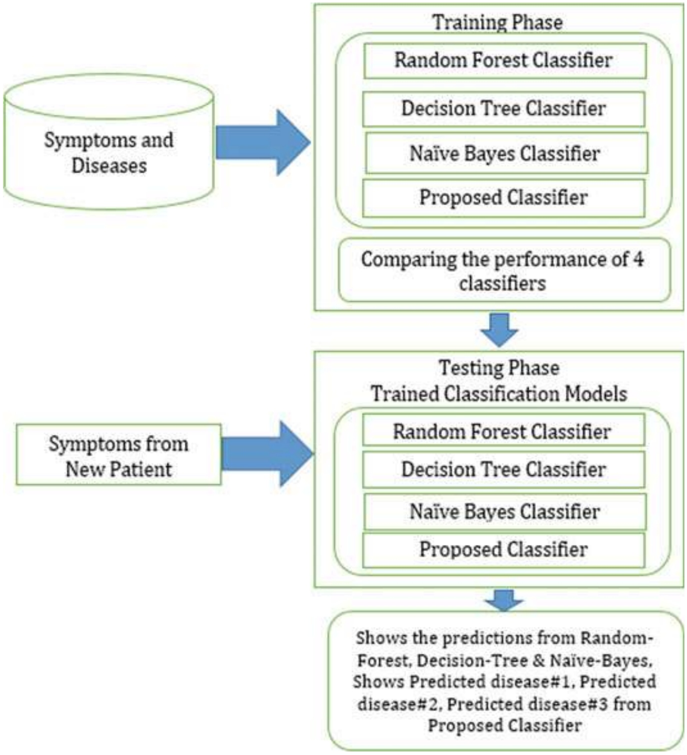
Patient Medical Records: The health history of the patient and diagnostics report can maintained in a single document called medical record or Electronic Health Records (EHRs) (Objective-3). It is readily available to both the patient and hospitals through electronic medical records.

Patient Surveys: This is the process for gathering the data from the various types of patients about their diseases, medical reports, treatment procedure, type of doctor, cost of the treatment, effect of the treatment, billing system, etc.

Comments from Individual Patients: Today social networking websites plays a vital role for gathering the opinion from various types of patients in healthcare industry. It gathers the comments from patients informally rather than by prepared questionnaire.

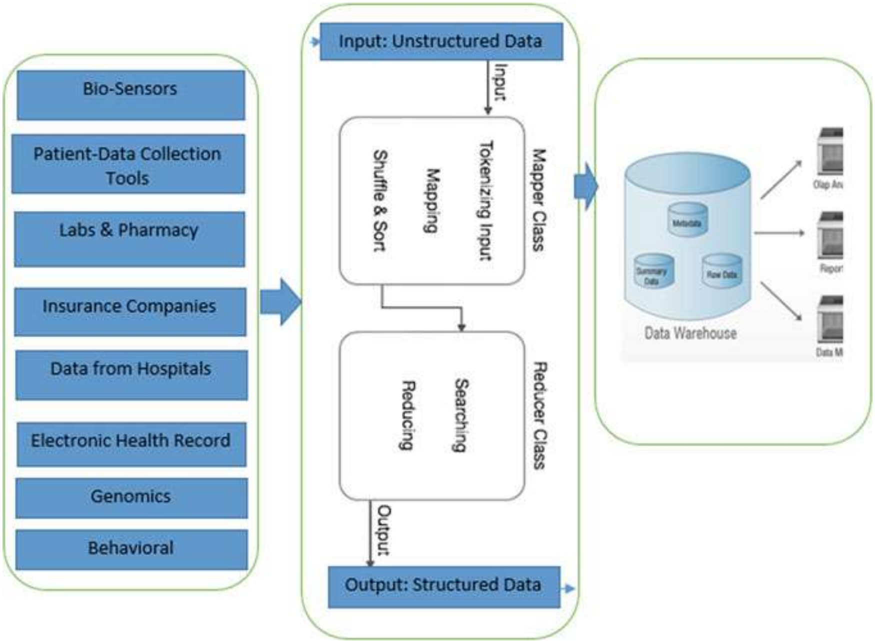
Standardized Clinical Data: The detailed information about each patient can be gathered from clinical and nursing homes, diagnostics centers and health agencies.

This concept was implemented through Python programming language and machine learning algorithms. The list of symptoms and diseases are stored in the form of dataset. The dataset contains 400 symptoms and 147 diseases which belongs various categories of diseases. Usually, the user can enter the list of symptoms in the system. Then, the system will finds the possible diseases as predictor-1, predictor-2 and predictor-3. The functionality of the proposed system is as follows:


Training Phase
Symptoms and Diseases
Random Forest Classifier
Decision Tree Classifier
Naïve Bayes Classifier
Proposed Classifier
Comparing the performance of 4 classifiers
Testing Phase
Trained Classification Models
Symptoms from New Patient
Random Forest Classifier
Decision Tree Classifier
Naïve Bayes Classifier
Proposed Classifier
Shows the predictions from Random-Forest, Decision-Tree & Naïve-Bayes, Shows Predicted disease#1, Predicted disease#2, Predicted disease#3 from Proposed Classifier

Method of Processing and Analysis

The data for healthcare domain is being generated from various internal and external sources in the world. The data can be gathered and processed in the following:



Web and social media data: The data from the social networking websites like Facebook, Twitter, LinkedIn, health plan websites and various apps.

Machine to machine data: Most of the unstructured data is being generated from sensors, meters and wearable devices.

Big transaction data: Patient join report and discharge reports, medical bills, healthcare claims, medical images are available in semi-structured and un-structured formats only.

Biometric data: Finger prints, genetics, handwriting, retinal scans, blood pressures, blood sugar, pulse and other personal details of each patient related to his body.

Clinical data: Semi-structured and unstructured data such as EMRs, physician's prescription, email, telemedicine details, etc.

After collecting the raw data, it can be stored in a data warehouse. Then, the big data analytics process the entire data and all types of data. Then, it handles various queries, it generates reports, OLAP and data mining. In big data era, many techniques and methods have been developed for aggregate, manipulate, analyze and visualize the healthcare data.

The healthcare big data can be handled by HADOOP from Apache. Hadoop is based on horizontal scalability large number of clusters of nodes, each node solve some part of the problem, integrates them for the final result.

4 Results and Discussion

The framework accepts the username and list of five symptoms. Then, it applied Decision Tree, Random Forest and Naïve Bayes classifiers on training and testing data. Finally, it gives predictions. Similarly, the proposed algorithm also works on training and testing data, finally it gives 3 predictions of diseases based on the given symptoms as shown in the following figures:

Table 1 shows the list of symptoms from the patient:

The above symptoms taken the machine learning classifiers as well as the proposed algorithm. Finally, it predicts and shows the possible diseases based on the given symptoms. The final results as shown in Table 2.

After giving the predictions, the accuracy has been calculated and shown in Table 3.

Among Decision Tree, Random Forest, Naïve Bayes and proposed algorithm, the accuracy is best in the proposed algorithm comparatively Decision Tree, Random Forest, Naïve Bayes.

Benefits.

The outcome of this paper is one type of software tool only. It collects and maintains very huge amount of data from various sources in the world related to health. It generates various types of reports and insights which gives precision treatment for the patients. The following are the expected benefits from this project:

- Medication is error-free
- Identification of high-risk patients easily
- It reduces hospital visits frequently
- It reduces patient waiting time in hospitals

Table 1 List of symptoms given by the patient

S. No	Symptoms				
	1	2	3	4	5
Input#1	Orthopnea	Fatigue	Dyspnea on exertion	Dyspnea	Shortness of breath
Input#2	Drowsiness	Sleepy	Pain chest	Angina pectoris	Pressure chest
Input#3	Wheezing	Cough	Shortness of breath	Chest tightness	Distress respiratory
Input#4	Hematuria	Tumor cell invasion	Pain	Anosmia	Thicken

Table 2 List of predictions given by Decision Tree, Random Forest, Naïve Bayes and proposed algorithm

S. No	Predictions					
	Decision Tree	Random Forest	Naïve Bayes	Proposed classifier		
				1	2	3
Input#1	Carcinoma of lung	Adenocarcinoma	Exanthema	Failure heart	Cardiomyopathy	Paroxysmal dyspnea
Input#2	Encephalopathy	Encephalopathy	Encephalopathy	Ischemia	Coronary arteriosclerosis	Coronary heart disease
Input#3	Exanthema	Sepsis (invertebrate)	Sickle cell anemia	Asthma	Chronic obstructive airway disease	Bronchitis
Input#4	Malignant tumor of colon	Pancreatitis	Encephalopathy	Neoplasm	Neoplasm metastasis	Carcinoma

Table 3 Accuracy of Decision Tree, Random Forest, Naïve Bayes and proposed algorithm

S. No	Decision Tree	Random Forest	Naïve Bayes	Proposed Algorithm
Input#1	0.8911564	0.9047619	0.9047619	0.976
Input#2	0.9047619	0.9047619	0.9047619	0.984
Input#3	0.8911564	0.9047619	0.9047619	0.986
Input#4	0.9047619	0.9047619	0.9047619	0.934

5 Conclusion

Each country taking more care about human health issues in today's world. Always WHO used to give so many suggestions for preventions of many epidemics or diseases. Today, the entire world is giving more importance to identification and prevention of many diseases based on the various symptoms from patients. Big data analytics plays an important role for getting predictions in healthcare industry. The healthcare industry is in a position to predict the disease based on given symptoms, and it will provide the suggestions healing of diseases. It reduces the patients to rejoin in the hospitals unnecessary. Physicians also get many suggestions about the good treatment for the patients. It provides the exact treatment for the patients and will provide the exact medicine. Automatically, it eliminates the side-effects for the patients. Finally, it helps the patients, doctors, hospitals.

References

1. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 33(7):1123–1131. https://doi.org/10.1377/hlthaff.2014.0041 PMID: 25006137
2. Dencelin LX, Ramkumar T (2016) Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures. Biomed Res S166–S173
3. Fang R, Pouyanfar S, Yang Y, Chen S-C, Iyengar S (2016) Computational health informatics in the Big Data Age: a survey. ACM Comput Surv 49:1–36. https://doi.org/10.1145/2932707
4. Balajee M, Suresh B, Suneetha M, Rani VV, Veerraju G (2010) Preemptive job scheduling with priorities and starvation cum congestion avoidance in clusters. In: 2010 second international conference on machine learning and computing, pp 255–259
5. Balajee M, Padmapriya G, Satish AR (2020) A framework for performance analysis on machine learning algorithms using covid-19 dataset. J Adv Mat: Sci J 9(10):8207–8215 (Publisher Advances in Mathematics: Scientific Journal)
6. Kaur P, Sharma M, Mittal M, Data B, Framework MLBSH (2018) Proc Comput Sci 132:1049–1059
7. Raghupathi, Raghupathi (2014) Big data analytics in healthcare: promise and potential. Health Information Sci Syst 2: 3
8. Sakr S, Elgammal A (2016) Towards a comprehensive data analytics framework for smart healthcare services. Big Data Res 4. https://doi.org/10.1016/j.bdr.2016.05.002
9. Sunil Kumar, Singh M (2019) Big Data analytics for healthcare industry: impact, applications, and tools. In: Big Data Mining and Analytics. ISSN: 222096-0654, 05/06, vol 2, no 1, pp 48–57, Mar 2019.https://doi.org/10.26599/BDMA.2018.9020031

View publication stats