A Framework for Disease Prediction Based on Symptoms using Big Data Analytics

Article · July 2021

CITATIONS
0

READS
179

1 author:

Kamal Gulati
Amity University
80 PUBLICATIONS **1,295** CITATIONS

SEE PROFILE

A Framework for Disease Prediction Based on Symptoms using Big Data Analytics

Dr. Kamal Gulati*

Abstract

The data is being collected from various sources in the world which belong to the 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 the technical market many BI is 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, Randomforest, Naïve Bayes and the proposed algorithm.

Keywords: Disease, Symptom, Decisiontree, Randomforest, 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 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.

In digital era, a very large amount of data is required for the healthcare industry. Now a days, many attackers are being tried to mine the sensitive data and release it to public domain. So data privacy plays a vital role in healthcare industry. Various methods and procedures are available for security and privacy of healthcare data. The characteristics of data security are as follows:

Authentication: It is the process of recognizing the user's identity. It checks the given credentials within an authentication server.

Encryption: It is the process to prevent unauthorized access of sensitive data. So the healthcare industry tries to provide confidentiality to the patient health records and it also provide privacy and security to the healthcare organization data.

* Associate Professor, Amity University, Noida, Uttar Pradesh, India, E-mail: drkamalgulati@gmail.com

Data Masking: It is the process to replaces the data which belongs to patient disease with an unidentifiable value. The data masking is one of the most popular approach to live data anonymization.

Access Control: It is the process of allocation of permission for accesses various files based on the need.

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 health care 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. Thisframework 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

The framework (Prableen Kaur et al., 2018) 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 data processing concept (Sunil Kumar et al., 2019) 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 centres and pharmacies and health insurance companies. This data can be handled by HADOOP framework.

Fang et al. (2016) proposed a framework titled "Health informatics processing pipeline framework" which consists of data capturing, storing, analysing, searching and decision support. It offers dynamic services to the patients through mobile devices and sensor networks. Legaz Garca et al. (2016) proposed a framework based on OWL. It gathers patient data (EHR) and utilized for data exploration.

Sakr and Elgammal (2016) proposed a method that integrates sensors, cloud, IoT and Big data analytics. It is able to handle patient profile analytics, population management etc. But not able to handle complex data sources such as images and streams. Pramanik et al. (2017) proposed a layered framework on healthcare system. This framework yields useful smart system services.

Dencelin and Ramkumar (2016) proposed a framework for analysing big data with the help of Apache Spark. It applies machine learning algorithms using different set of input features and network parameters. D.W. Bates et al.(2014), 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.

The framework which contains layers has been proposed for healthcare system by Raghupathi and Raghupathi (2014). 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. Theatrically these concepts are good enough.

The role of stakeholders in the framework for disease prevention are as follows:

Patients

Expecting customized healthcare service is very important for patient community. The patient community always tries to check the disease symptoms, side effects, experts' doctors, good hospitals, drug information etc. They also tries to follow the telemedicine when not able to go to hospital directly. The patient community always tries to know their health condition in every minutes by wearable devices.

Medical Practitioners

During diagnosis and treatment, massive amount of data is being generated like clinical services, lab results, medical images and sensor data. The data integration from various wearable devices, provides significant benefits which enhances the customized services to the patients.

Hospital Operators

Optimized facilities can be provided to the patients based on the locality and economical condition of the population in that locality. The big data analytics gives the exact count like how many number of staff are required based on the number of patients in the hospital.

Pharma and Clinical Researchers

The huge amount of big data which is being collected from various sources from healthcare industry is very useful to understand the biological and drug process with the help of predictive models. The information like drug recommended by a physician, opinion from patients on various drugs and its usage, sales history from drug shops etc.

Health Insurers

The big data analytics gives the information about the frequently occurring diseases. The insurance companies can introduce novel health plans with minimal premium cost. It recommends various plans to the customers based on the diseases.

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

Decision tree is very simple for gaining accurate and fast result, but the tree construction takes lot of time. It assigns class label to each patient for solving various problems. Random-Forest selects the most effective answer. It's appropriate for big knowledge bases and offers correct results by estimating missing data.Naive Bayes is the most simple algorithm that you can apply to your data. As the name suggests, Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors.

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 throughsurveys 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.

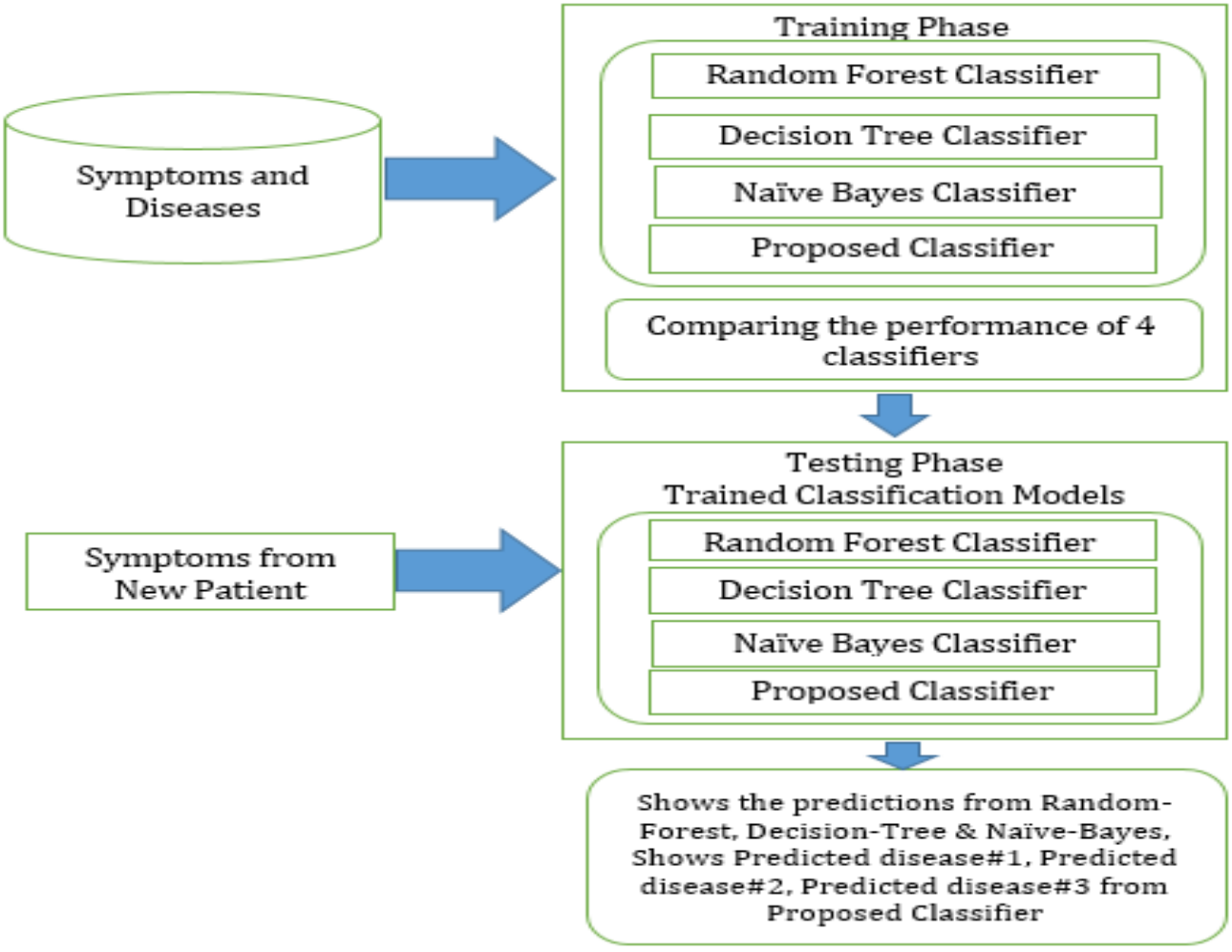
Patient Medical Records: The health history of the patient and diagnostics report can maintained in a single document called medical record or Electronic Health Records (EHRs) (Objective-3). It is readily available to both the patient and hospitals through electronic medical records.

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.

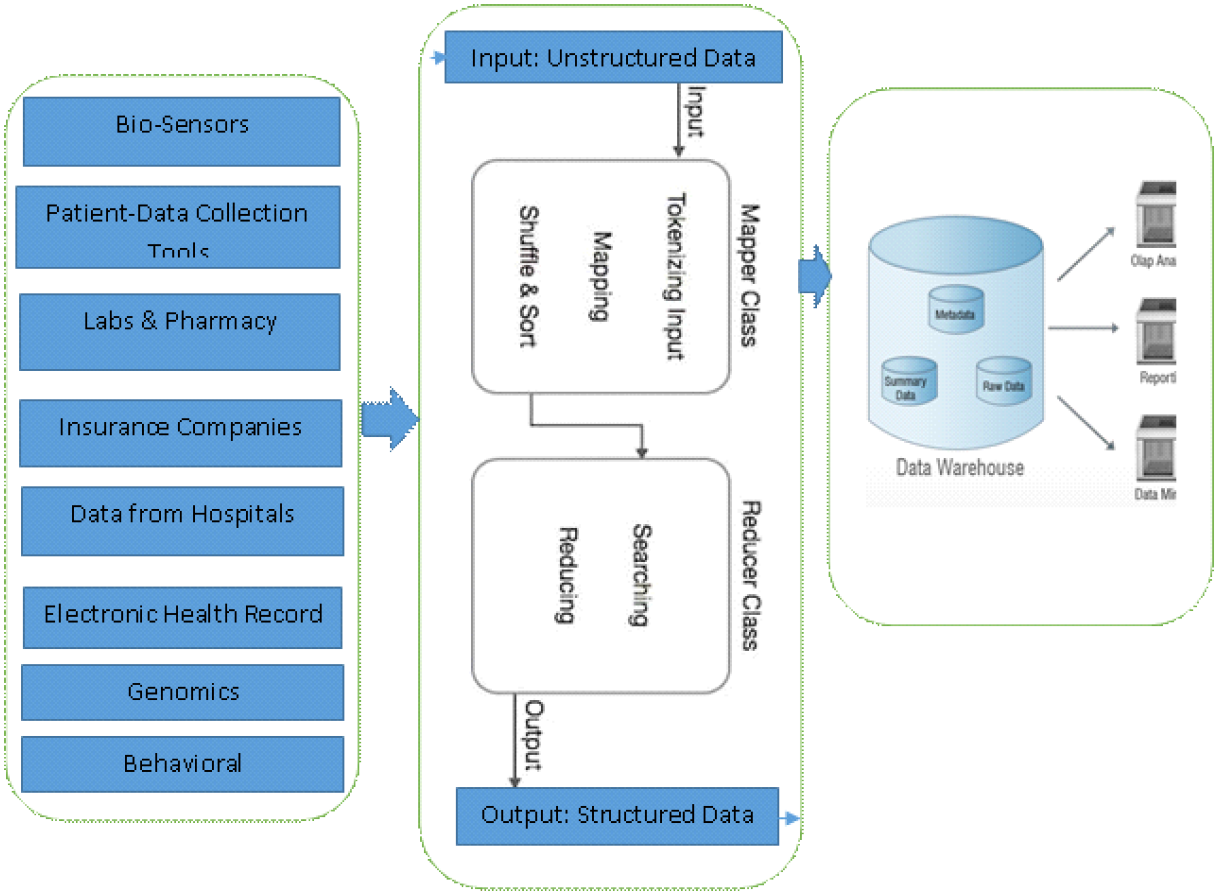
Standardized Clinical Data: The detailed information about each patient can be gathered from clinical and nursing homes, diagnostics centers and health agencies.

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:



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:



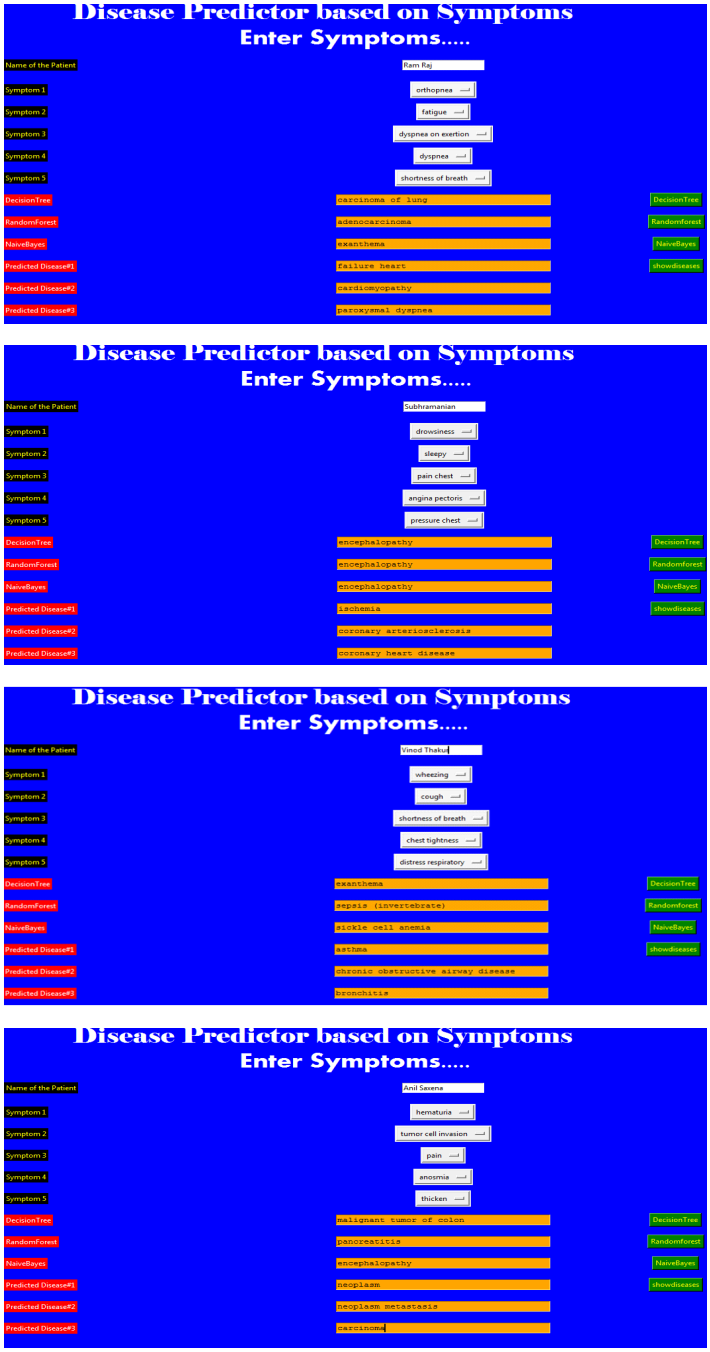
- Web and social media data: The data from the social networking websites like Facebook, Twitter, LinkedIn, health plan websites and various apps.
- Machine to machine data: Most of the unstructured data is being generated from sensors, meters and wearable devices.
- Big transaction data: Patient join report and discharge reports, medical bills, health care 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 open-source data processing platform called HADOOP from Apache. Hadoop is based on horizontal scalability and is able to process extremely large amounts of data by large number of clusters of nodes, each node solve some part of the problem, integrates them for the final result. This project can handles the issues related to healthcare industry including ownership, privacy, security and standards.

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:



The following table Table-1 shows the list of symptoms from the patient:

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

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 the following Table-2.

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

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

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

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

Limitations and Future Enhancements

Data Aggregation Challenges: The data related to healthcare system is being generated from various sources in the world like hospitals, administrative offices, Government offices, Private medical practitioners, Laboratories etc. Pulling it together and create a big data set is a challenge.

Polity and Process Challenge: After getting big data set, there is a need to protect it. This project can provide the protection through access control, authentication, encryption and decryption etc. But some part of the data is in the hands of cloud providers which are 3rd party vendors. So privacy and protection may be one of the challenge.

Management and Administration: It is one of the upcoming technology for the staff in healthcare industry. Again there is a need to recruit new IT experts for using HADOOP framework or need to give training to the existing staff. It increases the cost to the hospital authorities.

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. In Bigdata era, huge amount of data is being generated from various sources in the world. So Bigdata analytics plays an important role for getting predictions in health-care industry. The health-care 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. Prableen Kaur, Manik Sharma, Mamta Mittal,Big Data and Machine Learning Based Secure Healthcare Framework, Procedia Computer Science 132 (2018) 1049–1059.
2. Sunil Kumar, Maninder Singh, Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools, BIG DATA MINING AND ANALYTICS ISSN222096-0654, 05/06, pp48–57, Volume 2, Number 1, March 2019, DOI: 10.26599/BDMA.2018.9020031
3. Fang, Ruogu&Pouyanfar, Samira & Yang, Yimin & Chen, Shu-Ching & Iyengar, Sundararaj. (2016). Computational Health Informatics in the Big Data Age: A Survey. ACM Computing Surveys. 49. 1-36. 10.1145/2932707.
4. Sakr, Sherif&Elgammal, Amal. (2016). Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services. Big Data Research. 4. 10.1016/j.bdr.2016.05.002.
5. Raghupathi and Raghupathi: Big data analytics in healthcare: promise and potential. Health Information Science and Systems 2014 2:3.
6. Dencelin and Ramkumar, 2016, L.X. Dencelin, T. Ramkumar Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures, Biomed. Res., 2016 (2016), pp. S166-S173
7. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood). 2014 Jul;33(7):1123-31. doi: 10.1377/hlthaff.2014.0041. PMID: 25006137.
8. M Balajee, B Suresh, M Suneetha, VV Rani, G Veerraju, Premptive job scheduling with priorities and starvation cum congestion avoidance in clusters, 2010 Second International Conference on Machine Learning and Computing, 255-259
9. Dr. N Supriya Dr. Balajee Maram, Dr Satya Keerthi, Gorripati, A Framework Work For Data Security Using Cryptography And Video Steganography, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, 9 (4), 56-60
10. Maram Balajee, Padmapriya G.,Satish A.R., A framework for performance analysis on machine learning algorithms using covid-19 dataset, Journal Advances in Mathematics: Scientific Journal, 2020, Volume 9, Issue 10, pp.8207–8215, Publisher Advances in Mathematics: Scientific Journal.