ABSTRACT
Big data plays a key function in all aspects of organization especially in health science and clinical information. In the past days every health care related data contains data in the format of hard copy. (i.e.) Drug Prescriptions, reports, laboratory prescription etc. But, nowadays with the increasing need for ample variety of clinical data access over the internet, every data is move towards cloud in the outward appearance of digitization. The immense quantity of information collected from the various health cares, hospitals, and physicians etc. which are stored remotely on the cloud on top of the internet is called big data. This big data is in the appearance of structured (Tables in the outward appearance of rows and columns), semi-structured (XML data) and unstructured data (videos, X-rays, scan report, Images, Audios etc.). This enormous size of information needs to store, processed and retrieved in the effective manner to provide accurate and valuable information to the doctors, patients, house surgeon and medical researchers. This manuscript is mainly focused to predict and analyze the symptoms and also categorize the high-risk patients. Here the association rule is applied in the VPH (Virtual Physiological Human) method. VPH technique is used for collecting, analyzing, observing, categorizing and indexing the data which are observed from various hospitals, Patient record, primary health care, Insurance Firms etc. Here the experts from various fields are analyze and verify the data. Associate rule is used for uncovering the association among the symptoms in the database. From the data collected, we apply Apriori algorithm to the data base and also gives the risk priority category and highest risk patient.

Key words: Big data, VPH Technique, Associate Rule, Apriori Algorithm, HDFS, Map Reduce.

INTRODUCTION
Big data is the addition of vast quantity of information stored and processed on cloud by means of internet connection. Even a little quantity of data can also be represented as big data depending on the environment being used. This enormous quantity of data may be in the size of GB (Giga Bytes), TB (Terabytes), PB (Peta Bytes), EB (Exabyte). Multi-disciplinary information is processed in big data (Abinaya, 2015, Chitra et al., 2016) For e.g. Business, Media, Government Agencies and in addition to health care. Nowadays, Health care information has been dramatically increased which becomes additional difficult and expensive. But the present issue is to manage the huge number of information in the health science such as pharmaceutical information, clinical information, data from patient’s sensors in the hospital, Medical imaging such as X-Ray, Scan Report, Genomic Information, Historic Data of the patient, Public Health report, Government and private Hospital Health report, Disease Analysis about several diseases such as Respiratory disorder (Asthma), Diabetics, Heart Related diseases, etc. Electronic Health Record (EHR) is the term used in big data for health care (Sudharam et al., 2015, Haritha et al., 2015). In 2012, according to the report generated by the Common Wealth of Australia it has been analyzed that 500 Peta Bytes (PB) of data is all about Health care data. In 2020, this healthcare data will be estimated to 25,000 PB of data. The data which comes from U.S fitness care system alone reached 150 Exa bytes of data in 2011. In U.S every hospital system has the account of each patient history. Even if the patients switch physicians, the fitness care insurance firms 3 in the U.S have the proof of all the patients. In today’s world 90% of the information was created in the most modern years according to the data reported by Common Wealth of Australia in 2013. It has been predicted that the size of data will be generated in 2020 is 45 times superior than the present scenario. Every Day, 2.5 quintillion bytes of data has been generated. This brings about massive amount of high variety and velocity of data Kiyana et al., 2013; Matthew et al., 2013; Haritha et al., 2015; Dantanarayana et al, 2015; Nina et al., 2015; Harsh et al., 2014; Jasleen Kaur Bains, 2016)

CHARACTERISTICS OF BIG DATA
Volume, Velocity, Variety, Variability and Value are the 5 main characteristics of big data.
Volume: It represents the amount of data to be generated per second. Today, each society produce peta bytes of information each second.

PREDICTING THE HIGH-RISK PATIENT USING VIRTUAL PHYSIOLOGICAL HUMAN TECHNIQUE IN BIG DATA HEALTH CARE

A. Jameer Basha¹, S. Suguna², M. Kowsigan³, A. Sureshkumar⁴, V. Roopa⁵

Department of Information Technology, Sri Krishna College of Technology, Coimbatore – 641042, India.


¹jameer@skct.edu.in, suguna.s@skct.edu.in, kowsigan.m@skct.edu.in, a.sureshkumar@skct.edu.in,
²v.roopa@skct.edu.in
(Suguna et al., 2016; Kiyana et al., 2013; Matthew et al., 2013)

**Velocity:** It represents the rapidity of data to be delivered to the end users. Any information can be transferred and delivered globally within a fraction of second (David et al., Suguna et al., 2016; Kiyana et al., 2013; Matthew et al., 2013)

**Variety:** It represents the different kinds of data. In the past only structured data was used. These structured data can only be stored in the form of tables. But at the present all the structured and unstructured data (image, audio, video and sensors data) can also be stored and processed in the big data. (Kiyana et al., 2013; Matthew et al., 2013)

**Veracity:** It gives the correctness, reliability and consistency of data (Kiyana et al., 2013; Matthew et al., 2013)

**Value:** It represents the geometrical rate of data (Kiyana et al., 2013; Matthew et al., 2013; Suguna et al., 2016)

**BIG DATA ADVANTAGES**
1. Reduces the impact of death in advance.
2. Improve the quality and standard of patient’s life
3. Enhance the excellence of care and lower the cost
4. New medical researchers can learn and upgrade them easily.
5. Classify and visualize the disease easily.
6. Analyze the symptoms and improve the diagnosis easily.
7. It helps to provide medicine and dose correctly for the related diseases.

**ORGANISATION ROLE IN BIG DATA HEALTH CARE**

There are a bunch of organization and university have joined together to offer a solution for big data in health care. The University of Pittsburgh, Carnegie Mellon University (CMU) and University of Pittsburgh Medical Centre (UPMC) have joined together to make available a health package for analyzing and predicting diseases with the help of big data analytic tools. And the companies such as Apple and IBM collaborate together to introduce IBM’s Watson Health cloud service to achieve better treatment for the particular diseases such as Parkinson diseases, Diabetics, cardiovascular problems. Here, the iPhone and apple wearable watch user’s data will be collected through sensors placed on the mobile devices. For e.g. the person heart beat rate can be taken from the apple wearable watch. This data is directly stored on the IBM’s Watson health cloud. An Probability distribution based job scheduling had been performed in cloud (Kowsigan, et al., 2017).

**RELATED WORK**

Today many people in the earth suffers from lots of major health issues such as Cancer, Heart related diseases, asthma, viral fever, Diabetics and lung diseases etc.

![Figure 1. Big Data uniqueness](image)

**A. Big data part in health care for finding heart diseases?**

Today many people in the globe are affected by heart related diseases. Big data acts a most important role in order to save the patients’ health and to cut the death rate of heart patients. US-based Alivector Inc and the Apollo Hospital collaborated jointly to discover the Mobile ECG (Electro Cardio Gram) which examines the stroke and arrhythmia (irregular Heart Beat) screening all the way through mobile devices. The sensors which are embed on the mobile devices monitors the patient’s heart beat by simply rest it on their chest (Van-Dai et al., 2016). Heart disease
prediction has been performed by using fuzzy logic (Abhishek.T, 2013, Kowsigan, et al., 2017).

The patient health information is automatically recorded all the way through mobile phone devices in the form of ECG and then it is directly uploaded to the patient’s data bases. Here, all the data’s are gathered from the mobile devices, Heart Patients data collected from various hospitals, Data collected from experts for treating cardiac diseases, Clinical information, records in the paper format are transferred in to the digital format which is called HER (Haritha et al., 2016). The enormous quantity of information in the EHR is stored, processed and analyzed through big data by using HDFS & Map reduce. This data which is retrieved through big data analytics is helpful for patients, undergone training physicians, Doctors, Cardiac disease researchers etc. (Kiyana et al., 2013; Matthew et al., 2013; Suguna et al., 2016, Vaishal et al., 2016).

B. Big data part in health care for finding cancer tumors?
The American society of clinical oncology (ASCO) collaborates with SAP and they create Cancer Lin Q which really helps for oncologist to discover the appropriate tumor. The raw cancer data is collected from thousands of individual patients (who are affected from cancer) from various hospitals in U.S, Care Centre’s and health practices etc. This data is entered in to the computer system which is called Electronic health record (EHR). This EHR data is entered in to the knowledgeable data base on day by day basis. When this data base size increases we call it as big data. This big data which contains detailed information about cancer symptoms, identification of specific type of tumors, drugs for specific type of tumors, that could be helpful for future physicians, improve the cancer patient’s care, reduce the cost etc. Caner Lin Q runs on the SAP HANA platform. Where big data tools are applied to store, process and analyze the type of disease. (Hiba et al., 2015, Suguna et al., 2016)

C. Big data part in health care finding diabetics diseases?
Diabetic is then major disease affecting many people in the world. The ministry of Health in collaboration with the New Zealand society for the study of diabetics (NZSSD) have established the Virtual Diabetics Register (VDR). The VDR contains the detailed information collected from the various sources of laboratory orders for HbA1c(Measure of diabetics management), patient admission coded for diabetics in the hospital, Specific prescriptions of anti-diabetics therapies etc. The VDR is used to decide how many people are diagnosed with diabetics, their health conditions etc. So, in sequence to improve the diabetics prevention, today many applications have been created to monitor the blood sugar (glucose level), calories, Long term blood sugar (HbA1C), Total Triglycerides, Family history of high blood sugar, Waist measurement, height, Waist—to-hip ratio etc. The mobile apps used to
track diabetics are Diabetics tracker, Insulin pump timer, Diabetics pilot, Self Monitored Blood Glucose, Glooko, IBG Star, Glucose monitor, My Tele Care Diabetics etc (Noran and Al-Ardhi, 2017). These application data can be uploaded automatically to the respective server. So, all these information collected from the various sources are referred as big data. These big data are carefully processed with HADOOP YARN and HDFS to forecast the data accurately and improve the patient’s life and decrease the number of deaths.

**D. Big data part in health care finding Lung diseases?**

Chronic obstructive pulmonary disease (COPD) is the major disease corner to corner in the world. One among the four people is affected by COPD in the globe. Life Map solution is a company have Collaborated with Icahn School of Medicine at Mount Sinai– National Jewish Health Respiratory Institute founded a mobile application which is called COLD Navigator. It daily monitors then COPD patients Symptoms such as shortness of Breath, Phlegm are secreted by the mucus membrane during cold. Fatigue and chest tightness etc. It also has the capability to track the weather conditions, air quality alerts and other environmental factors which can cause and affect the COLD patients. If we record this information we have the options such as either save our data or directly send the e-mail to the physicians and care center teams. It also contains the digital guide book. This book offers the guidelines and methods to cope with a COPD diagnosis.

**PROPOSED WORK**

The information composed from the Mobile app, EMR, EHR, Clinical Data, wearable sensors, Genetic data, Primary rural health care, Insurance Claim are stored and processed with the help of extraction, Transform and Loading and also the data which is collected from the experts from popular Research universities are stored and processed in the secure manner with the help of HADOOP technologies (David et al., 2015; Chitra Pasupathi and Vijaya Kalavakonda, 2016). HADOOP is the technology used for storing and dealing out the massive quantity of data which may be in the form of structured (Relational data base) Semi-structured (XML data) and un-structured data (Video, image, Audio, G-mail, Word document, Graphics etc) (Poonam Vashisht & Vishal Gupta, 2016). This technology contains Map reduce and HDFS. Map reduce is used for processing the data and HDFS is used for storing the data. Performance analysis of wireless OCDMA systems using OOC, PC and EPC codes (Kanmani and Jameer Basha, 2016). After the processed data we apply VPH technique and association rule to predict and make the suggestions and decision for patient treatment (Marco Viceconti et al., 2015).

The VPH technique consist of three frames work i.e. Descriptive, Integrative and Predictive. Descriptive contains the data observed from various laboratories; hospitals in the globe are collected, categorized, shared and organized in a consistent manner. Integrative frame work contains the Experts from various disciplines to analyze the observations made in the descriptive frame work to provide multiple knowledge data base. Predictive frame work contains multiple methods to interconnect all the collected information from multi-disciplinary information to predict and analyze the disease. For finding the association amongst the symptoms, association rule is applied. Association rule are used in the knowledge base to forecast the symptoms and their respective diseases. Every data is written in the appearance of IF and THEN rules.

E.g.

\[
\text{IF 'Age of the patient' < 10} \\
\text{AND ‘the fever’ > 104 degree} \\
\text{THEN ‘Dolo’ is required.}
\]

IF --------> Information or facts which are called antecedent.

THEN -----> Conclusion or action that are Called consequent.

Multiple facts can also be joined by the keywords called AND (conjunction), OR (disjunction). This rule is used identify the association amongst the information in the knowledge base by using support and confidence and find the frequent IF and THEN patterns. Support is used to analyze how frequently the data item appears in the data base.

\[
\text{Sup (X) = No. of transactions which contain the Item set X / total no. of transactions}
\]

Confidence are used to identify the amount of times the IF and THEN statements found to be true.

\[
\text{Conf (X-> Y) = SUPP (X U Y)/ SUPP(X)}
\]

By using this association rule, we can easily recognize and forecast the high danger patients and symptoms linked with the disease.
APRIORI ALGORITHM PRINCIPLE

If the symptoms set in the data base is common, at that moment all of its subsets should also be common.

Consider the symptom set for skin diseases.

A = Skin razes, B = Skin cancer
C = Skin Aspirin, D = White Patches
E = Scalp Razes

The set may be in the form {A, B, C, D}, {A, B, C, D, E}{A, C, D, E}, {B, C, D, E}In the table the T.ID represents the Transaction Id.

The set may be in the form {A, B, C, D}, {A, B, C, D, E}, {B, C, D, E}, {A, C, D, E}In the table the T.ID represents the Transaction Id.

<table>
<thead>
<tr>
<th>T.ID</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The default minimum support count here is 4.

All the support count value in the table is larger than the minimum count. So, the table represents all are frequent data.

In the second step is to list all the 2-pairs of the frequent set.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>SUPPORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A, B}</td>
<td>4</td>
</tr>
<tr>
<td>{A, C}</td>
<td>5</td>
</tr>
<tr>
<td>{A, D}</td>
<td>5</td>
</tr>
<tr>
<td>{A, E}</td>
<td>3</td>
</tr>
<tr>
<td>{B, C}</td>
<td>6</td>
</tr>
<tr>
<td>{B, D}</td>
<td>5</td>
</tr>
<tr>
<td>{B, E}</td>
<td>4</td>
</tr>
<tr>
<td>{C, D}</td>
<td>7</td>
</tr>
<tr>
<td>{C, E}</td>
<td>6</td>
</tr>
<tr>
<td>{D, E}</td>
<td>4</td>
</tr>
</tbody>
</table>

In the third step is to list all the 3-triplets

<table>
<thead>
<tr>
<th>ITEM</th>
<th>SUPPORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A, C, D}</td>
<td>4</td>
</tr>
<tr>
<td>{B, C, D}</td>
<td>4</td>
</tr>
<tr>
<td>{B, C, E}</td>
<td>4</td>
</tr>
<tr>
<td>{B, D, E}</td>
<td>4</td>
</tr>
</tbody>
</table>

Hence, we compute the support count. Support count is for calculating the % of rule that satisfies.

SIMULATION AND RESULTS

In the initial step, the statistics collected from the various patients, hospitals, care centers, EMR (Kiyana et al., 2013), EHR insurance firms, genomic data (Priyanga et al., 2014) are entered in to the data base. The enormous degree of data is measured as big data. This big data is processed with the help of HADOOP ecosystem.

This big data is processed with HDFS and Map Reduce concepts. HDFS is used for storing the fitness care information and the Map reduce (Abinaya, 2015) is used for processing the medical data such as X-ray, ECG, Scan report, Pulse rate, heart beat rate etc (Vaishali et al., 2017). After performing the HDFS and Map reduce in the fitness care statistics we get the output as patient unique data.
In the next step, we estimate the overall count, support count and the confidence rate of the patient. This computation is based on the symptoms of a meticulous disease. Each numeric value represents the charge of the basic checkup result of each patient.

Figure 12: Patients Data record

In the third step, we extract the patients whose symptoms are matching. In the data base we take the Particular symptoms and count the overall patients who have the symptoms and we calculate the support and confidence in the database.

Figure 15: Calculate the support and confidence count.

In the fourth step, we compute the threshold limit of every patient by applying the attribute extraction and collecting the unique data. If the patient data exceeds the threshold limit at that moment we say that particular patient is in the highest risk.

Figure 16. Calculate the threshold limit
In the last step we compare the highly risky patients with the overall patients in the data base and also, we can check the patients who are in the high danger. So that we can take the remedial measures for the appropriate patients and also save their life in advance and improve the patient’s health.

CONCLUSION

Big data in fitness care acts a most important task in various hospitals, care centers, health care’s etc. The information collected from various sources are processed and analyzed with the help of HADOOP ecosystem. In the data base it contains the detailed information about each patient. Each patient has unique fields. In the fields they contain their information about BMI, Sugar level, Blood pressure etc. This data is applied with VPH technique with the aid of associative rule to uncover the high danger patient. This paper is focused to provide the care for the patients and develop their physical condition in the dangerous situations.

REFERENCES


Vaishali, G. and V. Kalaivani, Big Data Analysis for Heart Disease Detection System using Map Reduce Technique, International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE’16) Pp. 3-7 (2016).


Haritha Chennamsetty, S. Chalasani and D. Riley, Predictive Analytics on Electronic Health Records (EHRs) using Hadoop and Hive IEEE, International Conference on Electrical, Computer and Communication Technologies 10.1109/ICE CCT (2015)


Prabha S.M., Big data Solutions in Health Care: Problems and Perspectives, IEEE, Innovations in Embedded and Information, communication systems (2015).