I J C T A, 9(5), 2016, pp. 2153-2164

Disease Diagnosis for Personalised Health Care Using Map Reduce Technique

M. Mohan*, B. Vigneshwaran, G. Vineeth Raj and S. Harlin Jesuva Prince**

ABSTRACT

Analyzing enormous amounts of data available in biomedical, also known as “big data” is still increasing at a phenomenal rate. It provides an opening to develop a personalized healthcare with effective disease diagnosis and drug discovery using “Evidence based medicine” method. Recent advantages in biomedicine allows patient to get more access to their health information. The main aim is to make this information available to every individual preserving their privacy and by allowing the system to act as a “doctor” to diagnose the disease and provide the effective drug for it. The personalized healthcare system uses enormous amount of unstructured data gathered from various biomedical institutes and hospitals and using the map reduce technique the unstructured data is analyzed and based on the symptoms given by the patient the disease is effectively diagnosed and the best drug is identified. The process of collecting the data is done by the following methods 1.Analysis of patient health condition, 2.Formulating questions, 3.Evidence gathering and analysis, 4.Resultant output. As the volume of information fed increases the accuracy of prediction also increases and the existing information about the same disease is updated and if any new disease is identified it is added to the system. An automatic machine technique is used for disease discovery and its appropriate evidence based drug analysis is achieved.

Keywords: Big data, Disease diagnosis, Evidence based medicine, Map reduce, Hadoop.

1. INTRODUCTION

A survey taken by the Centers for Disease Control and Prevention revealed that about 610,000people die of heart disease in the United States every year, in 2005 survey, most defendants—92%—recognized pain in the chest as a symptom of a heart attack. Only 27% had knowledge of all major symptoms and knew to call 9-1-1 when someone was having a heart attack. About 47% of impulsive heart attack deaths occur outside a hospital. This recommends that many people with heart disease don’t act on initial warning signs. According to the International Diabetes Federation (IDF), about 371 million people worldwide were affected by diabetes, and 187 million of them had no idea that they have the disease. These statistics reveal the fact that there is still no effective disease diagnosis system available in the real time and most of the people were lazy and cannot afford to visit the hospitals for diagnosis. So, a Big Data driven approach towards individual health care, and its applicability to patient-centered outcomes, meaningful use, and reducing re-admission rates is necessary.

Medicine is driven by the single principle of having personalized medicine programs that will considerably improve patient care [6]. This can be accomplished by collecting medical history of patients worldwide and from that unstructured data the disease for the symptoms which are given as input can be effectively determined. The effective medication for the disease determined can be formulated by Evidence Based Medicine (EBM). Clinical practice guidelines are the embodiment of evidence-based medicine. Care organizations managed by the government began developing procedures in the 1990s to identify

* Research Scholar, Dept. of Computer Science & Engineering, Panimalar Engineering college, Chennai, India, *Email:* mohan.rm@gmail.com

** Student, Dept. of Computer Science & Engineering, Panimalar Engineering college, Chennai, India, *Email:* vvvicky440@gmail.com, gvineethrajjj@gmail.com and hjp2895@gmail.com

untimely medical care and reduce unnecessary utilization of services. More recently, policy makers in the state have integrated "best practices" or evidence-based guidelines in legislative proposals [2].

The information gathered from various sources should not reveal the patient details so patient privacy has to be a major concern. Improving the safety of the patients has become a chief attention of clinical care and research over the past two decades. A hospital's patient safety environment symbolizes a critical section of guaranteeing a safe environment and thereby can be more valuable to the prevention of adverse events [9].In this paper the system is allowed to act as a "doctor" in effective diagnosis of the disease by its symptoms and it helps every person to diagnose disease by himself it also lends a hand to doctors and researchers who are in need of disease information. Data in this system can me modified by authorized personnel only.

The remainder of the paper consists of the following: Section 2 presents some of the related works that are referred; Section 3 discusses the proposed work; Section 4 presents the results achieved along with discussions and Section 5 gives the conclusion along with limitations, followed by future work that could be added to the method.

2. RELATED WORKS

There had been numerous works in the area of disease diagnosis for individual health care is the past. There are multiple methods being stated differing. Mostly based on the technology used. But there is no actual implementation of the technique in wide level due to their respective limitation.

Big data is the new generation of technologies and architecture, designed to extract value from large volumes of a wide variety of data by enabling high-velocity capture, discovery and analysis. Using big data we can be able to retrieve using his medical reports. Below, we propose an overview of the systems using big data based approach and big-data with higher technology based approach.

2.1. Evidence Based System

Initially prototypes designed by using the evidence-based analysis using the big data technology on which the practice guidelines are used experimental practice guidelines are the embodiment of evident-based medicine. Practice guidelines specify the procedure of diagnosing and treating the particular conditions. This system uses various big data collection of the practice guidelines to provide an appropriate treatment for a particular. However there are gaps and inconsistencies in the medicinal literature associating one practice versus another as well as biased according to the perspective of the author [10].

2.2. Social Networks and Web Based Tools

Some technique use the social network approach using big data to be applied to life sciences in order to improve scientific and medical research. In which the social network sited such as Inspire, PatientsLikeMe, etc, are rich resources of patients' insights and clinical data from which the statistics can be processed to detect and provide diagnosis detail for the disease. One major limitation is that it take more time to process the transmission of information between the uses to technician (doctor) [7].

2.3. Psycho-Informatics Systems

Some other big data methods include the reliable measurements of emotion cognition and behavior study to predict the disease with many major advantage [3]. A most common related work which is in practice is, only machine learning method of patient clinical profile, that is available in existence but there is no processing of these data to detect the disease. Further the existing system studies only about the cognition and behavior using which we cannot diagnose the disease of the patient and when the disease is not diagnosed the effective drug for that disease cannot be found.

2.4. System Patientanomics

There are few methods which uses high technology along with big data but these methods needs advanced technology and costly to implement. One of these methods include the creation of virtual in-silico model to study the clinical and molecular science and to predict the medicine and disease of the prototype. This method has many limitation such as integrated analysis which will require extensive bioinformatics and biostatistics support. SAQ (safety attitudes questionnaire) methods are not suggested due to their limitation in number and ancient technology [5].

2.5. Search Engine Query Data

The search engine query data can also be used as a basis of big data. It is done by monitoring health-seeking behavior in the form of requests to search engines online, which are submitted by millions of users around the world each day. This method is to analyze a large numbers of Google search queries to monitor diseases like influenza in a population [2]. But the disadvantage of this method is that it relies entirely on the patient provided data which is highly unreliable and there can be a risk of fake information being provided. It provides information only about the influenza alone.

2.6. Behavior Studies

Behavior of a person on a great scale and on a fine detail can be studied simultaneously. Online services and universally possessed devices, such as modern cars, smartphones and other digital devices, keep track of our everyday activity. The result is an immense volume of unstructured data which offers numerous openings for tracking and analyzing behavior. Depression and excessive use of smart phones are the two methods which is used by this process to collect user data. There is no manual collection of data required, as it is directly available in electronic form [3]. This method involves the use of smart phones and other devices which may not be possessed by everyone and this again is an unreliable information. In addition to that there is a high chance of the data being misinterpreted or corrupted. So instead of using smartphones the clinical information of the patients can be used to account for the disease detection.

2.7. Omics and Clinical Health Data

Advances in various omics information are providing the footholds into establishing, for the first time, the genetic factors which are causal and that could help manage the following aspects of treatment: the right target, the right chemistry and the right patient. The challenge to be encountered by using this method is that funneling various biomedical data securely in to a unified system. Additionally, more studies will be needed to establish that personalized medicine and diagnostics accompanied by the use of computers directly benefit patients [4].

3. PROPOSED SYSTEM

The proposed system concentrates majorly on people who wants to know about their health status and diagnosis methods by sitting at their home and to avoid visiting the doctor and hospital. For concerns in existing systems an exact implementation of the system is not yet implemented, only the machine learning techniques are existing today.

The proposed work suggests a software application where the user can input the respective information i.e. the symptoms and get the corresponding diseases and diagnosis drug related to that symptoms.

This system uses k-means clustering algorithm. This data clustering method is one of the most widely used method where the datasets having “*n*” data points are divided into “*k*” groups or clusters. The definitive k-means algorithm works with in-memory information, but it can also be used for out-of-memory datasets.

The goal of K-Means algorithm is to find the best division of *n* entities in *k* groups, so that the total distance between the group's members and its corresponding centroid, representative of the group, is minimized. Formally, the goal is to partition the *n* entities into *k* sets S_i, $i = 1, 2, ..., k$ in order to minimize the within-cluster sum of squares (WCSS), defined as:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^j - c_j\right\|^2$$

where "*k*" is the number of clusters, "*n*" is the number of cases, "*c*" is the centroid for the respective cluster and $\left\|x_i^j - c_j\right\|$ provides the distance between an entity point and the cluster's centroid.

The K-Means is a greedy, computationally efficient technique, being the most popular representative-based clustering algorithm. The pseudocode of the K-Means algorithm is shown below.

```
Input:  $E = \{e_1, e_2, \ldots, e_n\}$(set of entities to be clustered)
        $K$(number of clusters)
        MaxIters(limit of iterations)
Output: $C = \{c_1, c_2, \ldots, c_k\}$ (set of cluster centroids)
        $L = \{l(e) \mid e = 1, 2, \ldots, n\}$ (set of cluster labels of $E$)
foreach $c_i$ € $C$ do
| $c_i \leftarrow e_j$ € $E$ (e.g. random selection)
end
foreach $e_i$ € $E$ do
| $l(e_i) \leftarrow$ argmin Distance $(e_1, c_j)$j € $\{1\ldots k\}$
end
changed $\leftarrow$ false;
iter $\leftarrow$ 0;
repeat
foreach $c_i$ € $C$ do
   | UpdateCluster($c_i$);
end
   foreach $e_i$ € $E$ do
      minDist $\leftarrow$ argminDistance $(e_i, c_j)$ $j$ € $\{1\ldots k\}$;
         if minDist$\neq l(e_i)$ then
            $l(e_i) \leftarrow$ minDist;
            changed $\leftarrow$ true
end
end
      iter++;
until changed = true and iter $\leq$ MaxIters;
```

The most common algorithm, described above, uses an iterative refinement approach, following these steps:

1. Define the initial groups centroids. This step can be done using different strategies. A very common one is to assign random values for the centroids of all groups. Another approach is to use the values of *K* different entities as being the centroids.
2. Assign each entity to the cluster that has the closest centroid. In order to find the cluster with the most similar centroid, the algorithm must calculate the distance between all the entities and each centroid.
3. Recalculate the values of the centroids. The values of the centroid's fields are updated, taken as the average of the values of the entities' attributes that are part of the cluster.
4. Repeat steps 2 and 3 iteratively until entities can no longer change groups.

An overall architecture of the proposed system using the k-means clustering is shown below:

The k-means architecture shows how to reduce the possible disease outcome into a finite and accurate set. Each Data Node is managed by their own Node Manager which computes the most resembling disease for the input symptoms. They internally use map-reduce technique and Machine learning to find the disease, from the given symptoms. The main purpose of k-means architecture is to forms a set of clusters from the given set of symptoms.

The proposed system has several modules. The detailed explanation of the system can be obtained from the modules namely:

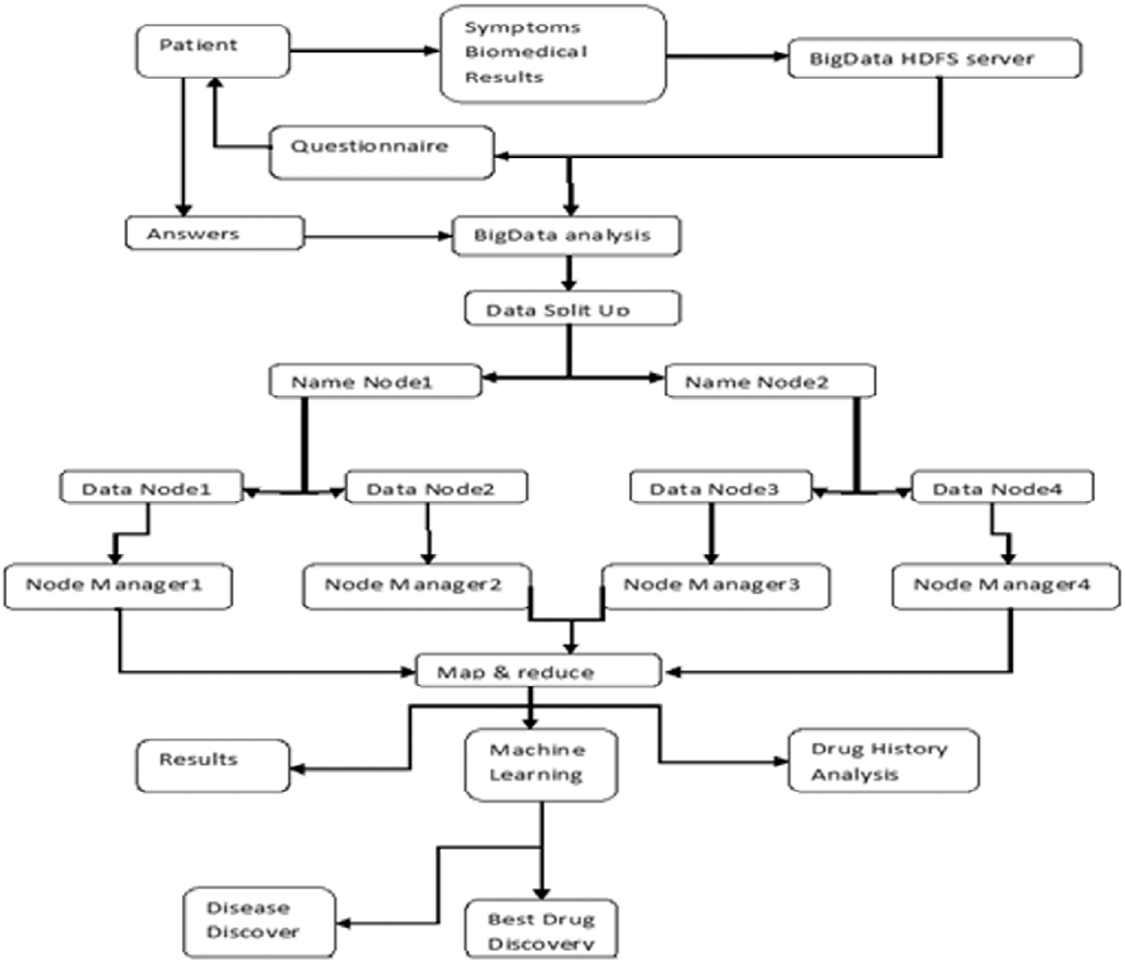

Figure 1: k-means architecture

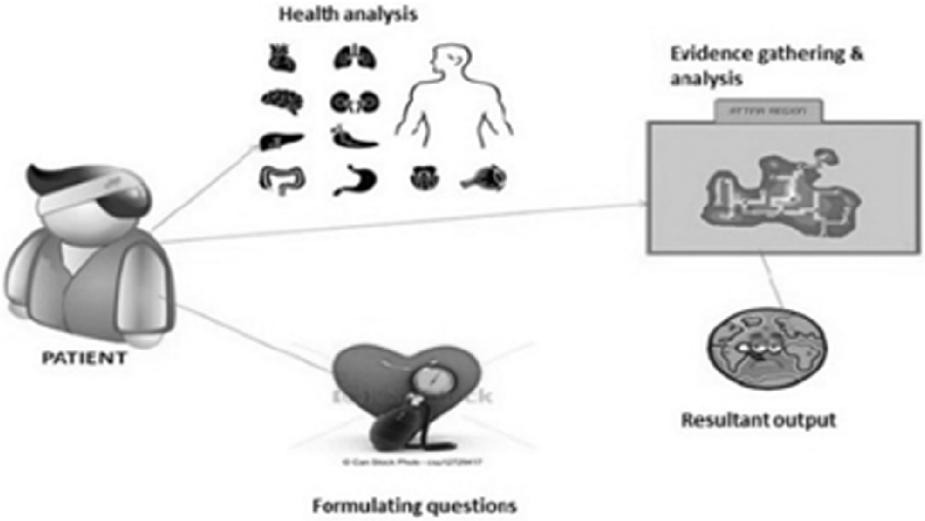

Figure 2: Proposed system architecture

1. Patient data gathering
2. Multi access controls
3. Patient accounts
4. Research account
5. Disease based data grouping
6. Machine learning algorithm

3.1. Patient Data Gathering

The patient who tries to use the system doctor is first prompted with a login page where the user has to sign in with his login and new users have to register their details for the first time. Now the patient details are being stored in the patient data base and the next time when he logins, the history of the previous diseases which the target patient have suffered will also be taken to consideration.

ALGORITHM

```
algorithm patientdata
begin
display login form to the user
get details about the user
submit details to sql query
if(username and password invalid)
then
    error(display(enter valid details))
end
else
    submit the form
end
```

3.2. Patient Accounts

Creating logins to user helps in maintaining the records of the patients who have registered in the system doctor. This patient accounts consists of details about patient's sugar level, glucose level, uric acid level, protein level, BP level. All these helps during data mining where the collective data of all the diseases, help to analyze the disease that have occurred to the patient.

ALGORITHM

```
algorithmpatientaccnt
begin
display the form to enter the details
glucose_level ← getParameter("glucoselevel");
sugar_level ← getParameter("sugar_evel");
protein_level ← getParameter("protein_level");
uricacid_level ← getParameter("uricacid_level");
update data to database
give unique id to the users
end
```

3.3. Formulating Questions

Questions are being formulated to get details from the general public about their view and experience about a particular disease and getting a collective data on them. These can be done is several ways. One may involve selecting people at random from a particular place and conducting surveys on the spot. Another may be a process in which people of a particular occupation or qualification are being selected and then survey is conducted viz., nurses and doctors in a particular specialization. In this method, forms are created which may be implemented online to get the feedback from the public.

ALGORITHM

```
algorithmformquesnre
begin
display the questionnaire
get the details about the person entering the survey
allocate priority based on the qualification of the individual
categorize opinions based on the type of disease
store the unstructured data on the distributed database
continue the steps for all the participants
end
```

3.4. Disease based data analysis

Disease based data analysis is the process of analysis or predicting the disease which the patient is having based on the details that are being collected from the user and submitting them to the unstructured data and using the data mining tools we sort out the most accurate disease which the patient may be probably having. Using map reduce technique we reduce the unrelated data items from the system and the most related disease are brought up in the cluster and produced as the output. Map reduce algorithm follow two stages: mapping phase, reducing phase.

ALGORITHM

```
algorithmdataanalysis
begin
mapping phase-
map ← declare a mapping function
give inputs in form of (key, value) pairs
loop: for each word equal to (key, pair)
increment counter
return the key value for the word ⇒ counter
reducing phase-
reduce ← declare a function to accept value from map
for each key-value pair
    add value to counter
return the word and counter as output
end
```

4. RESULTS AND DISCUSSION

The proposed system uses the big data map reduce technique to filter the data from unstructured data, it also uses the hospital provided data hence there is a minimum chance of error to occur.

In the disease based analysis the system uses the k-means clustering algorithm to group the datasets together. It can be shown in the diagrammatic form as:

When the patient gives 60 symptoms as input (say) for those 60 inputs the proposed system classifies them as 3 clusters (here) as shown in Fig 3. In each cluster the centroid is found so that the disease with

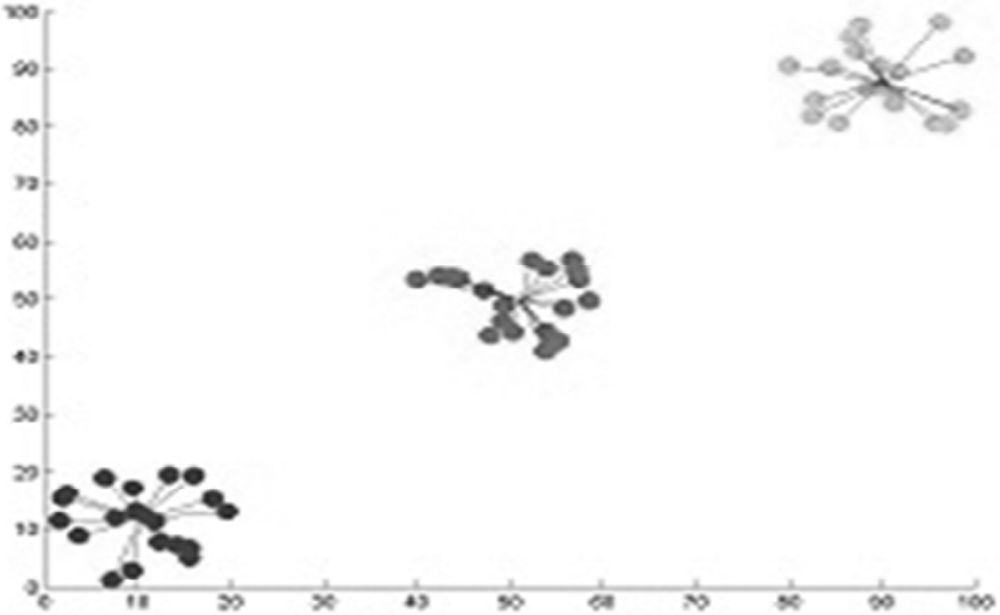

Figure 3: K-means for *N* = 60 and *C* = 3

resemblance symptoms can be found. The algorithm then greedily searches the cluster to locate the most appropriate disease. For every different "*N*" value we obtain different "*C*" value. If the number of "*C*" value is less then the disease proposed by the system will also be lesser than the relevant diseases available in the data.

The system uses machine learning technique in which the nonlinear unstructured data can be transformed to linear structured data.

Fig 4 shows the machine learning technique which is used to compute the non-linear separable function to a higher dimension linear separable function. If the user gives more number of inputs it is difficult to process it without machine learning so we use machine learning to reduce the search time required in the cluster.

Fig 5 depicts the relation between the number of clusters available and the searching time required for the algorithm for different "*N*" values. Considering a relatively small "*N*" value say N =3 we have small number of clusters and the searching time is also less but the efficiency of predicting the disease decreases as the value of "*N*" decreases. Now considering a relatively large "*N*" value say $N = 5$ or $N = 10$ the searching time exhaustively increases as the number of clusters but the efficiency of predicting the disease is much accurate when compared to a smaller "*N*" value. So it is always suggested that the system is provided with ample amount of symptoms (N) to get an efficient output.

Considering the "*N*" values sufficient enough to provide an accurate output if the clusters formed varies for each scenario then the accuracy increases as the number of clusters decreases. For example if the

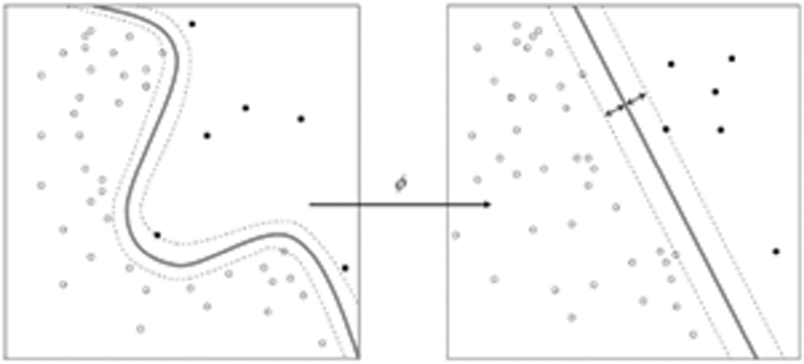

Figure 4: Machine learning

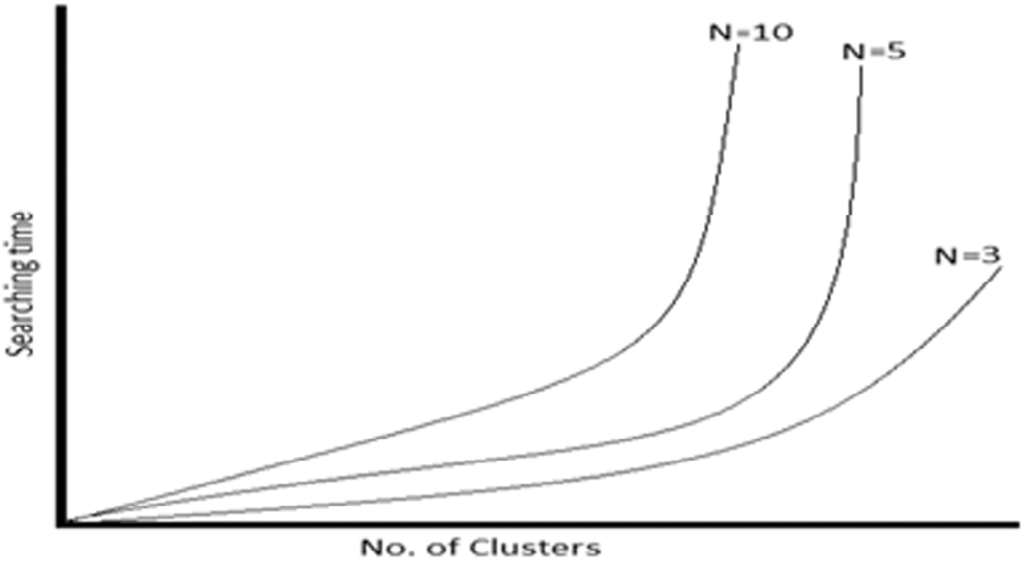

Figure 5: Efficiency in searching the cluster

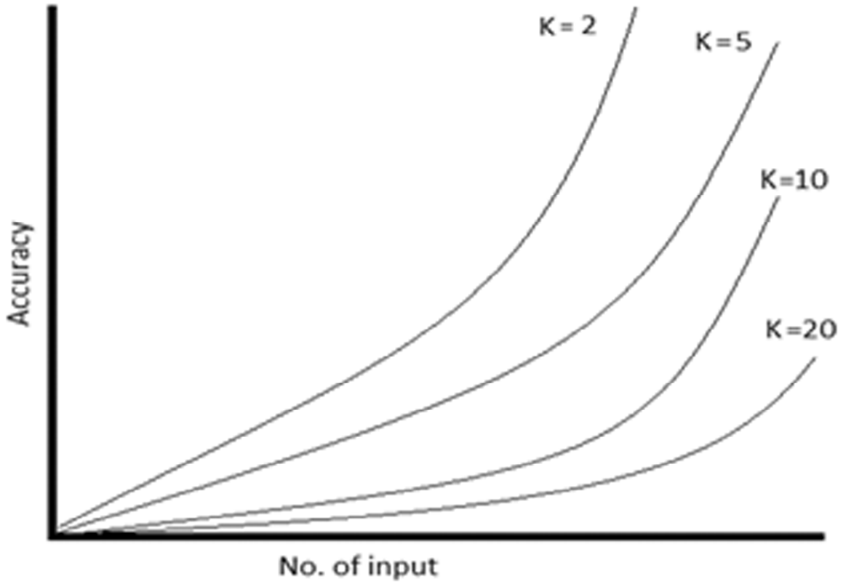

Figure 6: No of input vs Accuracy

Table 1
Illustration of system efficiency

Patient Input Symptom (#n)	*Relevant Disease count (#r)*	*System output (#o)*
3	5	4
10	20	3
8	15	6
5	13	3

number of clusters is 2 (say) then the accuracy of finding the effective disease is more when compared to a larger number of clusters say 20 as shown in Fig. 6. If the patient inputs relevant "*N*".

In the condition of patient giving 3 inputs (relatively small) the system should form small numbers of clusters instead, due to scarce in N, each separate symptoms forms as separate cluster and from each a relevant disease is yield. Here three clusters are formed and the possible disease count is 5, due to machine learning the system avoid one of the possible disease and predict only 4 output, which will be accurate without including the neglected one. Now assuming a larger input of 8, then the system now requires ample amount of time to separate those 8 symptoms into different clusters. But the strength of cluster would be comparatively less, meaning cluster sometime may hav much lower density than other clusters, in such cases those cluster can be avoided by performing the machine learning technique and only yield a most relevant disease, that is even though there are 15 possible outcomes the system ignore those with least possibility based on the inputted symptom and formed cluster. If the patients input a larger number of N and most relevant N, the system performance would increase and the number cluster forming can be effectively reduced, yielding faster performance.

In this system the user machine is used for auto diagnosis of the disease with reference to the user input of symptoms and reports. System will automatically identify the disease using machine learning algorithm. Server will store a set of trained dataset and its relevant diagnosis pattern. Using this algorithm disease are identified. Once the system identifies the user's input it is designed to represent data that accurately captures the state of the patient at all times. It provides a way to view the entire patient history not by using the previous medical records of the patients but keeping the entire patient information in a single file and assisting to ensure that the data is accurate, appropriate and legible. The chances of data replication is reduced as we use a single modifiable file which is updated constantly and eliminates the issue of lost forms or paperwork. This is the use of the evidence based medicine where the detection of the best drug is based on the patient history.

5. CONCLUSION

This paper describes the use of big data on disease diagnosis and drug prediction. The suggested method collects, stores, and analyzes massive amounts of indicative data at little cost and without risks or stress for patients or participants. Although sometimes overhyped, there is a great potential in the domain of biomedicine using computational approaches for the big data technologies in combination with other modelling strategies, and not in competition. This will minimize the risk of investments in research, and will ensure a constant improvement of in silico medicine, favoring its clinical adoption.

Though this system application provides most efficient medical data to the patients and doctors it still lacks in portability and offline access. Which are negligible on comparing with the benefits it can cause to the society. The future work considered includes plans to extend the system's capabilities by incorporating new services. These services include the following:

- Providing portability by extending the application to smart phones.
- Providing offline access so that the user can reduce the burden of using the internet.
- Integrating with social networks to collect more information and to benefit more users.

Although there are some modifications to the system there is still no practical implementation of the personalized health care system. Hence, this system provides an effective use to the doctors, researchers and patients.

REFERENCES

[1] Marco Viceconti, Peter Hunter, and Rod Hose, “Big data, big knowledge: big data for personalised healthcare”, DOI 10.1109/JBHI.2015.2406883, IEEE Journal of Biomedical and Health Informatics.

[2] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” Nature, vol. 457, no. 7232, pp. 1012-1014, 02/19/print, 2009.

[3] Alexander Markowetz, Konrad Błaszkiewicz, Christian Montag, Christina Switala, Thomas E. Schlaepfer,“Psycho-Informatics: Big Data shaping modern psychometrics”, Medical Hypotheses 82 (2014) 405-411.

[4] Scott J. Lusher, Ross McGuire, Rene C. van Schaik, C. David Nicholson and Jacob de Vlieg, “Data-driven medicinal chemistry in the era of big data”, Drug Discovery Today _Volume 00, Number 00 _ December 2013.

[5] D.V. Dimitrov, “Systems Patientomics: Thevirtual in-silicopatient”, New Horizonsin Translational Medicine 2(2014)1-4.

[6] Fabricio F. Costa, “Big data in biomedicine”, Drug Discovery Today Volume 00, Number 00, November 2013.

[7] Fabricio F. Costa, “Social networks, web-based tools and disease implications for biomedical research”, Drug Discovery Today Volume 18, Number 5/6 November 2013.

[8] Nitesh V. Chawla, and Darcy A. Davis, “Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework”, J. Gen Intern Med 28(Suppl 3):S660–5 DOI: 10.1007/s11606-013-2455-8.

[9] Natalie Zimmermann, Kaspar Küng, Susan M Sereika, Sandra Engberg, Bryan Sexton and René Schwendimann, “Assessing the safety attitudes questionnaire (SAQ), German language version in Swiss university hospitals-a validation study”, Zimmermann et al.. BMC Health Services Research 2013, 13: 347

[10] Twila Brase, R.N., President, “How Technocrats are Taking Over the Practice of Medicine”, Citizens' Council on Health Care, January 2005.

[11] Margaret A. Hamburg, and Francis S. Collins, “The Path to Personalized Medicine”, The New England Journal of Medicine, July 22, 2010.

[12] Alexandra Barratt, “Evidence Based Medicine and Shared Decision Making: The challenge of getting both evidence and preferences into health care”, Patient Education and Counseling 73 (2008) 407–412.

[13] Ruth Chadwick, “Ethical issues in personalized medicine”, Drug Discovery Today: Therapeutic Strategies, 2013.

[14] E. Elsebakhi, F. Lee, E. Schendel, A. Haque, N. Kathireason, T. Pathare, N. Syed, R. Al-Ali, “Large-Scale Machine Learning based on Functional Networks for Biomedical Big Data with High Performance Computing Platforms”, http://dx.doi.org/doi:10.1016/j.jocs.2015.09.008.

[15] Zhuyuan Fang, Xiaowei Fan, Gong Chen, “A study on specialist or special disease clinics based on big data”, DOI 10.1007/s11684-014-0356-9.

[16] S. Suganya, “Big care: the Solution for Doctors to Identify Disease Using Big Data”, Transactions on Engineering and Sciences, Special Issue on International Conference on Synergistic Evolutions in Engineering (ICSEE) – 2015.

[17] Wullianallur Raghupathi and Viju Raghupathi, “Big data analytics in healthcare: promise and potential”, Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3.