DISEASE PREDICTION USING MACHINE LEARNING OVER BIG DATA

Vinitha S, Sweetlin S, Vinusha H and Sajini S

Computer Science and Engineering, S.A. Engineering College, India

ABSTRACT

Due to big data progress in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care and community services. When the quality of medical data is incomplete the exactness of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may results in weakening the prediction of disease outbreaks. In the proposed system, it provides machine learning algorithms for effective prediction of various disease occurrences in disease-frequent societies. It experiment the altered estimate models over real-life hospital data collected. To overcome the difficulty of incomplete data, it use a latent factor model to rebuild the missing data. It experiment on a regional chronic illness of cerebral infarction. Using structured and unstructured data from hospital it use Machine Learning Decision Tree algorithm and Map Reduce algorithm. To the best of our knowledge in the area of medical big data analytics none of the existing work focused on both data types. Compared to several typical estimate algorithms, the calculation exactness of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

KEYWORDS

Big data analytics, machine learning, healthcare.

1. INTRODUCTION

With the advance of big data analytics equipment, more devotion has been paid to disease expectation from the perception of big data inquiry, various explores have been conducted by choosing the features mechanically from a large number of data to improve the truth of menace classification rather than the formerly selected physiognomies. However, those prevailing work mostly measured structured data. Thus, risk organization based on big data analysis, the following tasks remain: How should the mislaid data be lectured? How should the main chronic diseases in a positive county and the main faces of the disease in the region be gritty? How can big data analysis expertise be used to estimate the disease and generate a better method?

To solve these problems, it see the structured and unstructured data in healthcare field to assess the risk of disease. First, the system use Decision tree map algorithm to generate the pattern and causes of disease. It clearly shows the diseases and sub diseases. Second, by using Map Reduce algorithm for partitioning the data such that a query will be analyzed only in a specific partition, which will increase the operational efficiency but reduce query retrieval time. Map reducing algorithm is used for partitioning the medical data based on the output of Decision Tree map

DOI:10.5121/cseij.2018.8101

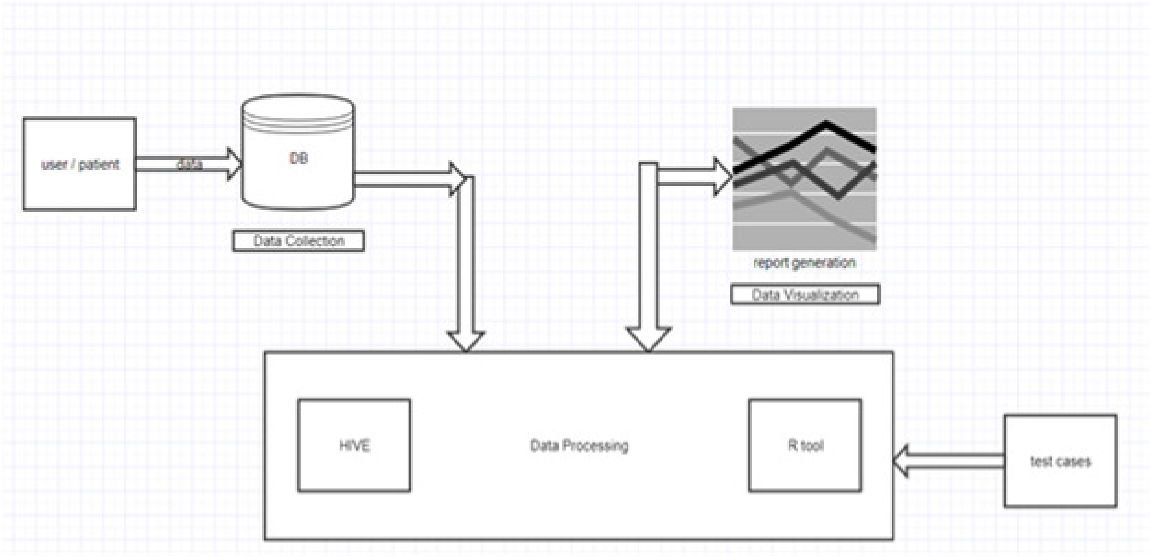
algorithm. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm increases.

1.1 OBJECTIVE

The analysis accuracy is reduced when the quality of medical data in incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. However, those existing work mostly considered structured data. There is no proper methods to handle semi structured and unstructured. The proposed system will consider both structured and unstructured data. The analysis accuracy is increased by using Machine Learning algorithm and Map Reduce algorithm.

2. DETAILED DESCRIPTION

2.1 ARCHITECTURE DIAGRAM



2.2 MACHINE LEARNING

Machine Learning (ML) delivers methodologies, approaches, and apparatuses that can help resolving analytic and predictive hitches in a miscellany of medicinal areas. ML is being used for the inquiry of the wild of controlled edges and their mixtures for forecast, e.g. forecast of illness development, removal of medicinal information for consequence investigation, treatment guidance and provision, and for the overall enduring organization. ML is also being used for statistics examination, such as discovery of proportions in the data by rightly commerce with flawed data, clarification of incessant data used in the Strenuous Care Unit, and brainy troubling subsequent in real and ordered nursing. It is contended that the successful presentation of ML attitudes can help the tally of computer-based structures in the healthcare setting providing chances to ease and enhance the exertion of medical boffins and eventually to recover the competence and excellence of medicinal repair. Below, it précis some main ML requests in medicine. Machine Learning learns the data and produces the result.

2.3 MAP REDUCE

Map Reduce is a essential constituent of the Apache Hadoop software plan. Hadoop allows hardy, feast dispensation of huge shapeless facts sets crosswise product processor bunches, cutting-edge which each bulge of the bunch covers its own packing. Map Reduce assists two crucial tasks: It tracts out slog to innumerable nodes within the group or map, and it classifies and reduces the consequences from each node into a consistent response to a inquiry.

2.4 REPORT

A boom producer is a processor package whose determination is to income data from a source such as a database, XML stream or a spreadsheet, and use it to produce a article in a arrangement which contents a specific human circulation. Here the system is making increasing boom of patient.

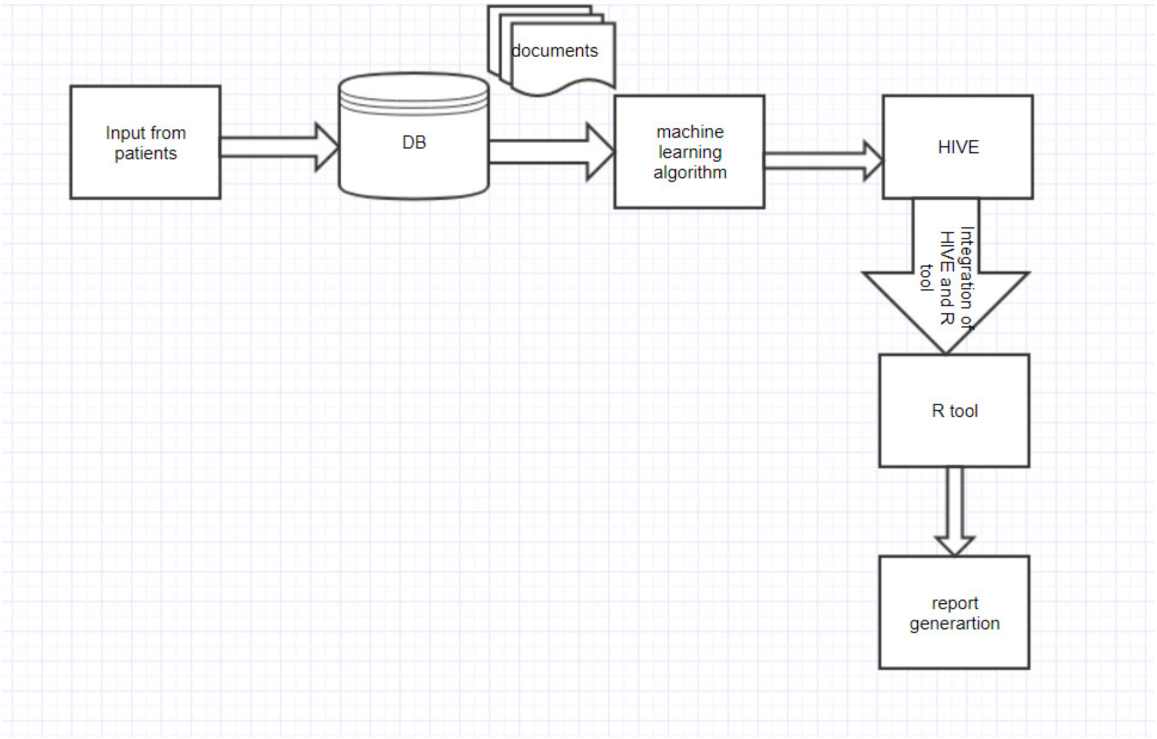
2.5 EXISTING SYSTEM

Machine can predict diseases but cannot predict the sub types of the diseases caused by occurrence of one disease. It fails to predict all possible conditions of the people. Existing system handles only structured data. The prediction system are broad and ambiguous. In current past, countless disease estimate classifications have been advanced and in procedure. The standing organizations arrange a blend of machine learning algorithms which are judiciously exact in envisaging diseases. However the restraint with the prevailing systems are speckled. First, the prevailing systems are dearer only rich people could pay for to such calculation systems. And also, when it comes to folks, it becomes even higher. Second, the guess systems are non-specific and indefinite so far. So that, a machine can envisage a positive disease but cannot expect the sub types of the diseases and diseases caused by the existence of one bug. For occurrence, if a group of people are foreseen with Diabetes, doubtless some of them might have complex risk for Heart viruses due to the actuality of Diabetes. The remaining schemes fail to foretell all possible surroundings of the tolerant.

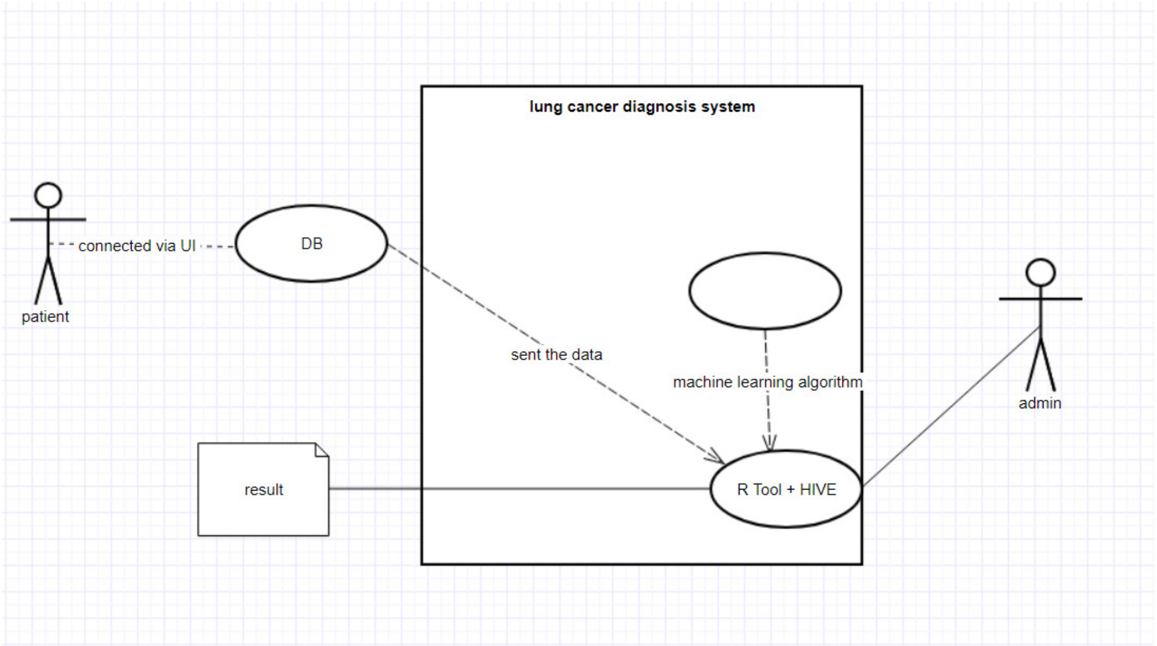
2.6 PROPOSED SYSTEM

Our application will be at affordable cost. Decision Tree Machine Learning Algorithm predicts Diseases as well as all sub diseases. Map Reduce Algorithm is implemented to increase operational efficiency. It reduces Query retrieval time. Accuracy is improved using Machine Learning algorithm. The proposed system begin with the thought that was not executed by the ancestors. It gadget Decision Tree machine learning procedure for calculating diseases as well as calculating all the other thinkable sub diseases. It member Map Reduce algorithm for subdividing the data such that a request would be scrutinized only in the explicit partition, which will increase effective proficiency but cut query rescue time. In tally to that, it provide definite rations for specific clients to pattern his/her condition. Thus making our presentation broadly open by all at cheap cost.

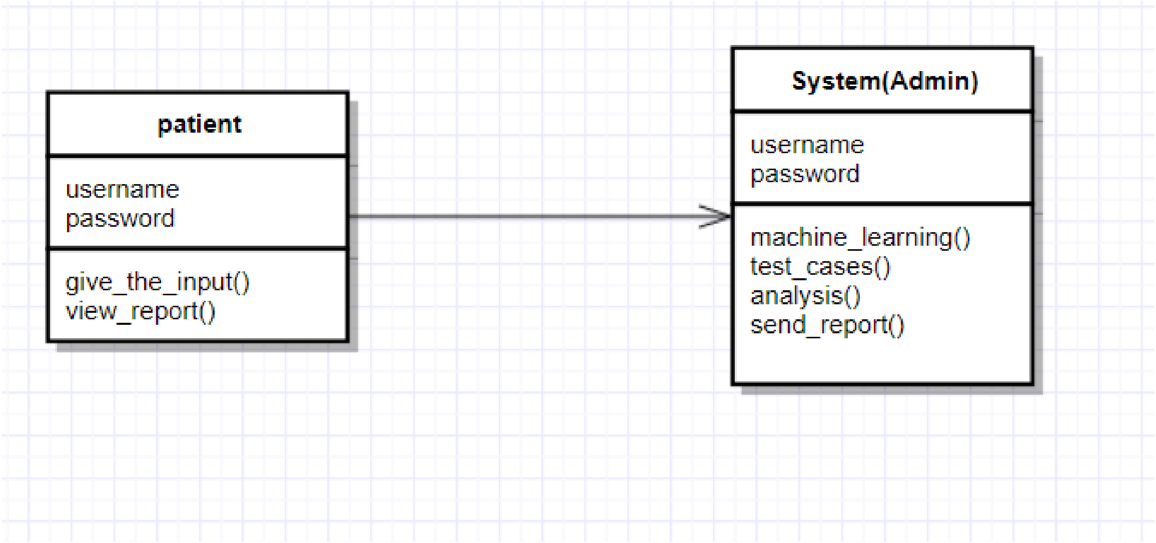
2.7 FLOW CHART



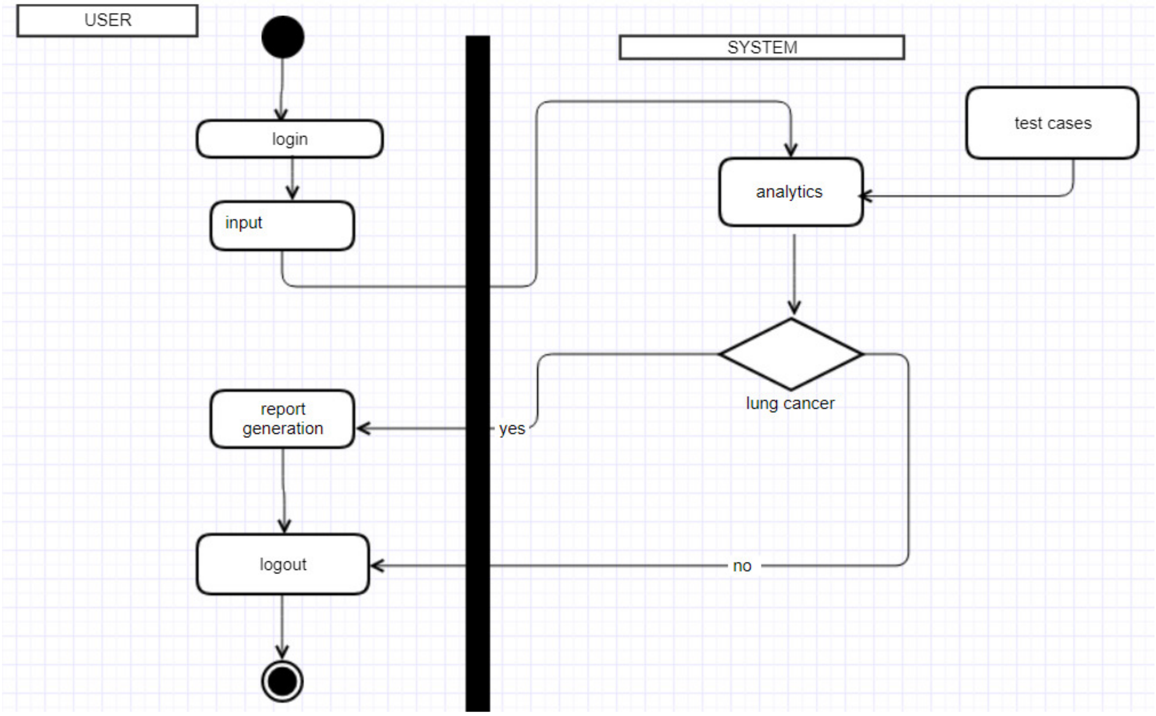
2.8 USECASE DIAGRAM



2.9 CLASS DIAGRAM



2.10 ACTIVITY DIAGRAM



3. RELATED WORK

[1] In 2010, Apache Hadoop sharp big data as "datasets which could not be apprehended, succeeded, and managed by general computers within an okay scope." On the basis of this definition, in May 2011, McKinsey & Company, a global accessing help said Big Data as the next edge for improvement, war, and yield. Big data shall callous such datasets which could not

be attained, succeeded and stored by standard database software. This classification includes two associations: First, datasets dimensions that obey to the usual of big data are shifting, and may cultivate over time or with scientific developments. Second, datasets measurements that adapt to the ordinary of big data in unalike submissions contrast from each other.

[2] Clinical data recounting the phenotypes and dealing of patients denotes an underused data font that has much bigger research likely than is currently grasped. Mining of electrical health records has the facility to form a new patient-stratification doctrines and for tight fitting unknown disease links. Mixing EHR data with genetic data will also give a more kind of genotype-phenotype affairs. However, a wide series of permitted, ethical, and methodological reasons presently hold back the organized confession of these data in electrical health histories and their excavating. Here, it consider the likely for furthering medical examination and experimental care using EHR data and the tasks that must be dazed before this is a truth.

[3] The medical resources of many countries are limited. For example, in China, the growth of medical resources is not balanced that 80% people are living in areas with inadequate medical resources while 80% medical resources are allocated at the big cities. Construction of big health application system by successfully mixing medical health resources using smart depots, health Internet of Things (IoT), big data and cloud computing is the vital way to resolve the above difficulties. Big health is a talented industry, which is characterized by people-center, managing a person's health from birth to decease, from anticipation to rehabilitation and involving industry from administration to market. The field of big health covers health goods field (including the drugs, medical devices, elder goods), health service field (including medical services, income services, mobile healthcare), fitness real estate field (including pension, healthcare) and health finance field (including health protection and other financial products).

[4] Chinese herbal products (CHPs) are commonly developed for patients with hyperlipidemia in traditional Chinese medicine (TCM). Since hyperlipidemia and connected sickness are public topics worldwide, this training discovered the drug shapes and occurrences of CHPs for giving patients with hyperlipidemia. Traditional Chinese medicine (TCM) has become common as a healing for central indicators in patients with hyperlipidemia. This drill likely to study the treatment patterns of TCM for patients with hyperlipidemia. The study population was recruited from a random-sampled troop of 1,000,000 folks from the National Wellbeing Insurance Exploration Record between. It recognized 30,784 fatality visits linked with hyperlipidemia judgment and collected these medical records. Overtone rules of facts withdrawal were led to moveable the co-prescription plans for Chinese herbal products (CHPs).

[5] In this paper, it witness the use of recurrent neural networks (RNNs) with the situation of search-based operational publicity. It practice RNNs to map equally queries and ads to real valued vectors, by means of which the significance of a given (query, ad) couple can be simply calculated. On upper of the recurrent neural networks, it familiarize a novel consideration network, which studies to assign attention scores to different word locations according to their intent importance (hence the name Deep Intent). Later by this method, the path output of a arrangement is computed by a weighted sum of the hidden states of the RNN at each word according their attention scores. The system achieve end-to-end exercise of together the RNN and attention system below the guidance of user click logs. These worker click logs are sampled from a commercial search engine. It demonstrate that in most cases the attention network improves the quality of learned vector representations, evaluated by AUC on a physically labeled dataset. And furthermore, it highlight the effectiveness of the learned attention nicks from two aspects as:

query rewriting and a modified BM25 metric. The system illustrate that using the learned attention scores, one will be able to produce sub-queries that would be of better qualities than those of the state-of-the-art methods. In count, by regulating the term occurrence with the care scores in a normal BM25 formula, one is bright to improve its performance evaluated by AUC.

[6] Abstract Traditional wearable devices have various drawbacks, such as uncomfortableness for long-term wearing, and insufficient accuracy, etc. Thus, health monitoring through traditional wearable devices is hard to be sustainable. In order to obtain and manage healthcare big data by sustainable health nursing, the system design “Smart Clothing”, enabling unobtrusive collection of various physiological indicators of human body. To offer persistent cleverness for smart clothing erection, mobile healthcare cloud stand is constructed by the usage of mobile internet, cloud computing and big data analytics. This paper announces design facts, key tools and applied implementation methods of smart dress system. Typical claims powered by smart clothing and big data clouds are presented, such as medical backup response, emotion care, disease diagnosis, and real-time tangible interaction.

[7] In this it extant a new deep learning manner Bi-CNN-MI for paraphrase identification (PI). Created on the vision that PI needs associating two sentences on many heights of granularity, it learn multigranular decree images using convolutional neural network (CNN) and model boundary features at each level. These topographies are then the input to a logistic classifier for PI. All limits of the model (for embeddings, convolution and classification) are straight optimized for PI. To address the lack of training data, the system pretrain the network in a novel method using a language modeling task. Results on the MSRP corpus surpass that of earlier NN competitors.

[8]Does the estimate of lung cancer using the double dispensation system. The image dispensation system is familiarized into the double for early prophecy. The challenging in this progression is recognition of tiny nodes which comprehends early cancer finding. The unstipulated knobs in lungs can be spotted using ridge recognition algorithm.

[9] It proposed a system that integrates different datum such as gene information, DNA methylation, and miRNA. In this paper, the model has combined multiple kernel learning methods and dimensionality reduction.

[10] On the available data mining algorithms to classify the data and extract the knowledge from it. It discusses about the difficulties in classification, segmentation, extraction and selection. It compares the different algorithms like Support Vector Machine, Naïve Bayesian classification, Rough set theory, Decision Tree.

[11] Does the study on decision trees and their behavior in arriving to the conclusion. Tree node splitting based on relevant feature selection is a key step of decision tree learning, at the same time being their major shortcoming: the recursive nodes partitioning leads to geometric reduction of data quantity in the leaf nodes, which causes an excessive model complexity and data over fitting. In this paper, the author presented a novel architecture called a Decision Stream.

[12] The paper is based on the smoking behavior of the user. The e-cigarette has a small electrical resistance coiled wire in 1.5 ohms which is connected to the positive and negative poles of the device. When the button of e-cigarette is pressed, the resistance coil can be connected with electrical supply under the immersion of some “E-liquid”, the coil heats up and transform the E-

liquid to vapor, which can be inhaled by the smokers. It monitors the smoking behavior of the user in order to prevent the patient from cancer.

4. CONCLUSION

In this paper, it bid a Machine learning Decision tree map algorithm by using structured and unstructured data from hospital. It also uses Map Reduce algorithm for partitioning the data. To the highest of gen, none of the current work attentive on together data types in the zone of remedial big data analytics. Compared to several typical calculating algorithms, the scheming accuracy of our proposed algorithm reaches 94.8% with an regular speed which is quicker than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm and produces report. The report consists of possibility of occurrences of diseases.

REFERENCES

[1] "M. Chen, S. Mao and Y. Liu. Big data: A survey".

[2] "P. B. Jensen, L. J. Jensen and S. Brunak. Mining electronic health records: Towards better research applications and clinical care".

[3] "Yulei wang1, Jun yang2, Viming.Big Health Application System based on Health Internet of Things and Big Data".

[4] "S.-M. Chu,W.-T. Shih,Y.-H. Yang, P.-C. Chen and Y.-H. Chu. Use of traditional Chinese medicine in patients with hyperlipidemia: A population-based study in Taiwan".

[5] "S. Zhai, Chang, R. Zhang and Z. M. Zhang. Deepintent: Learning attentions for online advertising with recurrent neural networks".

[6] "M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu. Smart clothing: Connecting human with clouds and big data for sustainable health monitoring".

[7] "W. Yin and H. Schutze. Convolutional neural network for paraphrase identification".

[8] "Weixing Wang and Shuguang Wu. A Study on Lung Cancer Detection by Image Processing".

[9] "Thanh Trung Giang, Thanh Phuong Nguyen and Dang Hung Tran. Stratifying Cancer Patients based on Multiple Kernel Learning and Dimensionality Reduction, 2017 IEEE 9th International Conference on Knowledge and Systems Engineering (KSE)".

[10] "Saranya P and Satheeskumar B. A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 713-719".

[11] "Dmitry Ignatov and Andrey Ignatov. Decision Stream: Cultivating Deep Decision Trees", 3 Sep 2017 IEEE".

[12] "Kelvin KF Tsoi1, Yong-Hong Kuo and Helen M. Meng. A Data Capturing Platform in the Cloud for Behavioral Analysis Among Smokers An Application Platform for Public Health Research", 2015 IEEE".