International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 3 47 - 50

An Efficient Medical Text Mining in Diseases Diagnoses And its equivalent Data privacy Preservation Policy for Medical Data: A Review

Miss. Kalyani P. Barabde
Department of Computer Science and Engineering
G.H. Raisoni Collage of Engineering and Management
Amravati, India
barabdeka@gmail.com

Prof. Nitin R. Chopde
Department of Computer Science and Engineering
G.H. Raisoni Collage of Engineering and Management
Amravati, India
nitin.chopde@raisoni.net

***Abstract*—** Healthcare systems use a medical text mining which have been increasingly facilitating health condition monitoring and disease modeling. System works on the Personal Health Information (PHI) of the user. Healthcare system grant users access to range of health information and medical knowledge. Benefit of the system is all the information about disease, precautions and healthcare are store at one place. Unfortunately, delegating both storage and computation to the untreated entity would bring a series of security and privacy issues. One of the controversial issues for PHI is how the technology could threaten the privacy of patient health information. The proposed system focused on fine-grained privacy-preserving static medical text access and analysis, which can hardly afford the dynamic health condition fluctuation.

Keywords- *Differential Diagnosis, Image Processing, fuzzy Logic, K-means Clustering, Image segmentation, Pattern Reorganization, fuzzy pattern.*

*****

I. INTRODUCTION

In a Human Life diseases are major cause of illness and death in the modern society. Medical diagnosis is an important task that should be performed accurately and efficiently and its automation would be very useful. Due to increased computing, doctors have always made use of technology to get help in various possible ways, from surgical imagery to X-ray photography. Unfortunately, technology has always stayed back when it came to diagnosis, process that still requires a doctor's knowledge and experience to process the large number of variables involved, ranging from medical history to nature conditions, and various other factors. The number of variables counts up to the total variables that are required to understand the complete working of nature itself, which no model has successfully analyzed yet. To get out of this problem, medical decision support systems[1] are becoming more essential, which will assist the doctors in taking correct decisions. Medical decision is a highly specialized and challenging job, especially in case where diseases show similar symptoms, or in case of rare diseases. Unfortunately all doctors are not equally skilled in every sub specialty and they are in many places a scarce resource. A system like automated medical diagnosis would enhance medical care and reduce costs. Some conventional algorithm overlook various factors involved such as prevailing conditions, the build-ups resulting in the symptoms, family history, medical history and other factors relating to the patient, due to sheer magnitude of available unknown variables. Normally experienced doctors classify diseases based on the different diagnosis[2] method. This involves narrowing down the diseases to the root disease out of the list of diseases which shows similar symptoms. This is done using their knowledge and experience, and it is then confirmed by performing various tests. Especially in some areas, the problem of lack of trained and experienced doctors leads to intensification of this problem[3]. So we are trying to build this process of differential diagnosis to make this rather tough task a lot easier. The method is further modified to use the images to collect the symptoms by processing that images and then reduce the number of underlying variables to only one variable by finding the root disease, using smart pattern matching involving *k*-NN classification technique[4] and the next probable diseases by performing differential diagnosis. Using all these, and by database having a medical history of the user, the probability of disease occurrence may get calculated, despite of the various unknown variables. The system will output the disease from the symptoms entered by the user and also gives the next highly probable disease, and thus, the most effective course of action to be performed can be determined. System works on the Personal Health Information (PHI) of the user. Unfortunately, privacy and security issues have significantly impeded the wide adoption of healthcare systems, since the physical health information disclosure and mistreatment would bring about extremely serious privacy leakage for the patients.

The system, using various techniques mentioned, will in twist display the root disease along with the set of most likely diseases which have similar symptoms. This system will give the list of diseases that the patient has maximum probability of suffering from. This, in turn, will help to recommend specific tests corresponding to diseases in the list, thus reducing the number of non-consequential tests and thus resulting in saving time and money for both the doctor and the patient.

II. LITERATURE SURVEY

Basically, we can say that the medical diagnosis process can be interpreted as a decision making process, throughout which the physician induces the analysis of a new unknown case from an available set of medical data and from her/his clinical experience. At the University of Calabria in Italy, the medical decision making process has been cornputeriized, Physicians at the Cosenza General Hospital currently arc using the diagnostic decision support system to help them with the timely identification of breast cancer in patients through The application of a well-defined set of classification data. Dr. MirnmoConforti presented the system in 1999 & he explained the architecture fom this particular point of view, emphasizing the powerful efficiency and effectiveness of Mathematical Programming approaches as the basic tools for the design of the Computer Aided Medical Diagnosis system. MimmoCnnforti addresses our attention to early detection of cancer on the basis of small amount of clinical information. Hubert Kordylewski[5], Daniel Graupe[6] in 2001 describes the application and principal of a large memory storage and retrieval (LAMSTAR) neural network. The LAMSTAR was specifically useful for application to problems having very large

memory that contains many different categories or attributes, like where some of the data is exact while other data are fuzzy and where, for a givenproblem, there may be some data categories are totally missing.The LAMSTAR network is fast and can shrink/grow in dimensionality without any reprogramming. LAMESTER network is a self-organized map (SOM) with link weight between two neurons of this SOM module. The network also having features of forgetting and of interpolation and extrapolation, thus being able to handle partial data sets. Applications of the network to three specific medical diagnosis problems are described: two from nephrology and one related to an emergency-room drug identification problem.

DejanDinevski, Peter Kokol, GregorStiglic, Petra Povalej[7] elaborates the use of self-organization to combine different specialist's opinions generated by different intelligent classifier systems with a purpose to raise classification accuracy. Early and accurate diagnosing of various diseases has proved to be of vital importance in many health care processes. In recent years intelligent systems have been often used for decision support and classification in many scientific and engineering disciplines including health care. However, in many cases the proposed treatment or the prediction or diagnose can vary from one intelligent system to another, similar to the real world where different specialists may have different opinions. The main aim here is to imitate this situation in the manner to combine different opinions generated by diverse intelligent systems using the self-organizing abilities of cellular automata. Because most ensembles are construct using definite machine learning method or a combination of that method, but the drawback being this is that the selection of the appropriate method or the combination of that method for a specific problem must be made by the user. So, to overcome this problem anensemble of classifiers is constructed by a self-organizing system applying cellular automata (CA). Jenn-Lung Su, Guo-Zhen Wu[8] introduced the concept of database has been widely used in medical information system for processing large volumes of data. Author says that numeric and symbolic data will define the need for new data analysis techniques and tools for knowledge discovery. In his paper three popular algorithms for data mining which includes Decision Tree (DT), Bayesian Network (BN), and Back Propagation Neural Network (BPN) were evaluated. The result shows that Bayesian Network had a good presentation in diagnosis ability.

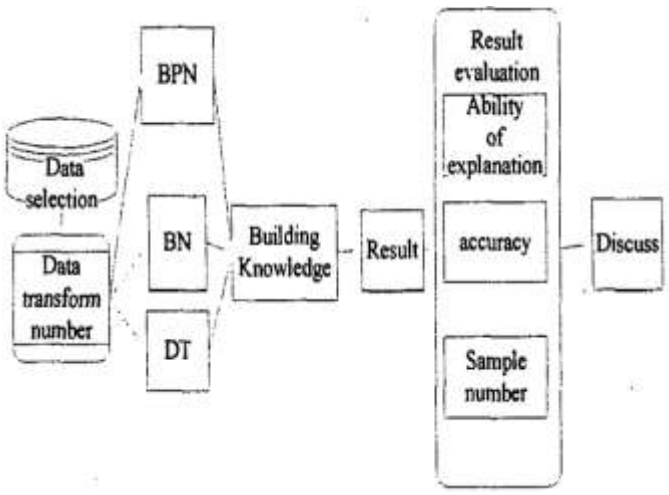

Figure 1: Procedure of Knowledge Discovery[8]

Rebeck Carvalho, Amiya Kumar Tripath, Rahul Isola[9]introduced the concept of Medi-Query. Paper says that traditionally the huge quantities of medical data are utilized only for clinical and short term use. Medi-Query gives idea to use this huge storage of information.so that diagnosis using this historical data can be made. There are systems to predict diseases of the heart, lungs, and brain based on past collected data from the patients. Paper focus on computing the possibility of occurrence of a particular ailment from the medical data by mining it using algorithm which boost accuracy of diagnosis by combining Neural Networks, Bayesian Classification and Differential Diagnosis all incorporated into one single approach. The system uses a Service Oriented Architecture (SOA) wherein the system components of diagnosis, information portal and other miscellaneous services provided are coupled. It will also help the medical society in the long run by helping them in getting accurate diagnosis and sharing of medical practices which will facilitate faster research and save many lives. Shucheng Yu, Cong Wang, KuiRen, Wenjing Lou in their paper "Attribute based data sharing with attribute revocation" [10] addressed an important subject of attribute revocation for attribute based systems. In particular, they considered practical application scenarios in which semi- trustable proxy servers are available, and proposed a scheme supporting attribute revocation. The Ciphertext-policy attribute based encryption (CP-ABE) was used to propose the system. CP-ABE is a public-key cryptography primitive that was proposedto resolution the exact issue of fine-grained access control onshared data in one-to-many communications.

Fully homomorphic encryption (FHE) is also widely studied and obviously provides a believable solution to secure outsourcing computation. However, some intrinsically unsolvable problems significantly obstruct its wide application in practice. Most existing work is mainly based on the polynomially bounded hard problems in lattice and the plaintext has to be encrypted bit-by-bit. Recently, Jung et al. proposed a privacy preserving data aggregation supporting multivariate polynomial evaluation without secure communication channel, respectively in one aggregator model and participant's only model. However, when it is applied to out sourced medical text mining, it only suggests static statistics computation, leaving the patient's dynamic health condition monitoring that can more precisely reflect her/his suffering status untouched. Moreover, the addition aggregation and multiplication aggregation are achieved in independent mechanisms, which lead an additional load on power-restricted users. Hsu et al. proposed an image feature extraction in encrypted domain with privacy-preserving scale-invariant feature transform (SIFT)[11], by exploiting Paillier's cryptosystem. However, it cannot be used in outsourced medical image feature extraction.

Healthcare systems have been increasingly facilitating health condition monitoring, using medical text mining and image feature extraction. The paper by Jun Zhou[12] gives a privacy-preserving dynamic medical text mining and image feature extraction scheme PPDM. Here author talks about the security and privacy issues of the user's Personal Health Information (PHI). Therefore, he designs a secure and privacy preserving outsourcing medical text mining with image feature extraction. By using one way trapdoor function, a fully homomorphic data aggregation is conduct which shows the basis for proposed privacy-preserving protocol for dynamic medical text mining. Next, an outsourced disease modeling and early interference is achieved, respectively by devising a privacy preserving function correlation matching from dynamic medical text mining and designing image feature extraction.

III. PROPOSED WORK

Training Phase

In training phase we create a database by applying fuzzy rules on the various symptoms collected by doctors. Our proposed methodologies may contain 50 to 100 symptoms of various diseases. The symptoms by the doctor are entered manually as well as we are collecting the symptoms from images.

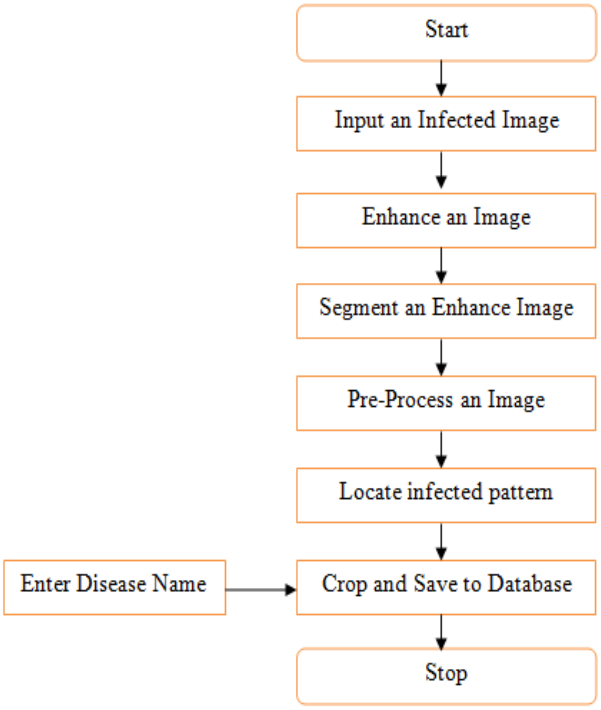

Figure 2: Training Phase for image input

We are using an Image Enhancement process in which noisy pixels of an image get filtered this process can be achieved with gabor wavelet filter and we can express it as

Fimage= $\int_1^{h*w}$ [(pr)(pb)(pg)]*Ef

Where,

pr,pg,pb=Image Pixel components

Ef=Enhancement Factor

h=Height of an Image

W=width of an Image

Fimage=Filter Image

When an image is filtered and noisy pixels are removed, future step is to segment that image either with canny edge detection or sobel edge detection method. For each pixel line image segmentation can be expressed as,

$0< |P_{Back}–P_{forward}|<=10$

Set $P_{forward}$=255(for first hit only). For second, third and onward set P_{Back}=255

An image pre-processing in a proposed method is a step where an infected pattern is located in an enhance image. Processing is a step where an extreme level pixel are identified and expressed differently for gray scale images and RGB images.k-Means clustering algorithm is used to cluster disease patterns where all patterns are gathered around it's reference disease name. All patterns are clustered as per Ecludian distance measure and fall into multiple clusters.

Testing Phase

In Testing phase we apply fuzzy rules on symptoms taken by the doctors then Pre-process it &take out one conclusion. Testing phase may work with the following steps:

1. Collect symptoms from patients.
2. Pre-process symptoms.
3. Extract features from database, apply fuzzy rules &out one proper conclusion.
4. If conclusion out is about more than one diseases then proposed methods needs more symptoms &processed from step (2).
5. This method will be iterative from step (2) to (4) until it does not output a single and accurate

Disease

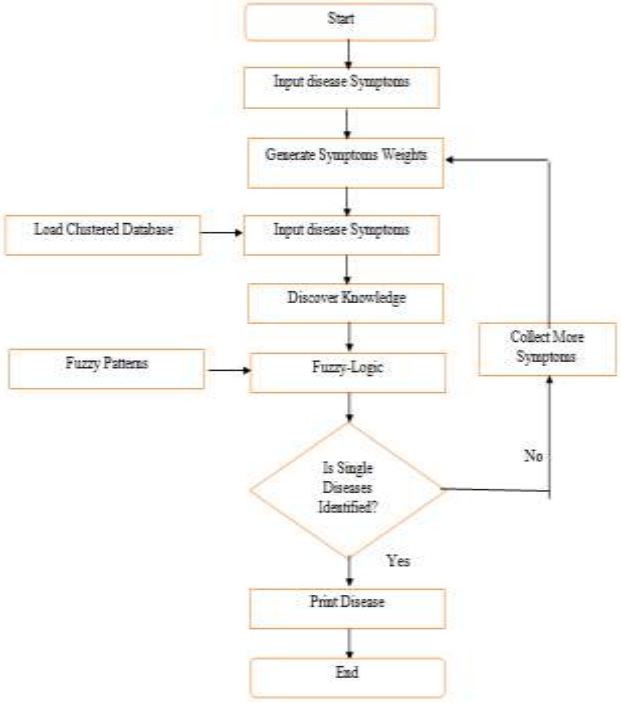

Figure 3: Proposed System DFD

Fuzzy patterns are used in the situation when there is difficult to conclude if multiple patterns are input which may fall into more than one disease, decision will taken by fuzzy rules Patterns. MDFC means multiple disease fuzzy conclusion. If multiple diseases are found with similar ranking, it becomes difficult to pinpoint to one of them, when no more symptom is unique to any single disease affecting its ranking. In such cases, recent medical historical data stored in the database of the proposed system is used, with a time period of three months to rank the diseases on basis of the probability of their occurrence in the review period. Time frame of three months provided accurate diagnosis with less error. The duration depends mainly on the seasonal cycles in the location where the system is to be implemented. If the interpreted diagnosis is still vague, then the system proceeds to differential diagnosis for the differential diagnosis, weights are assign to each disease whenever it is entered in the system. According to the weights the decision can be taken out.

REFERENCES

[1] R. A. Miller, "Medical diagnostic decision support systems—Past, present, and future—A threaded bibliography and brief commentary," J. Amer. Med. Inf. Assoc., vol. 1, pp. 8–27, 1994

[2] W. Siegenthaler, Differential Diagnosis in Internal Medicine: From Symptomto Diagnosis. New York: Thieme Medical Publishers, 2011.
[3] S. F.Murray and S. C. Pearson, "Maternity referral systems in developing countries :Current knowledge and future research needs, "Social Sci. Med., vol. 62, no. 9, pp. 2205–2215, May 2006
[4] L. Li, L. Jing, and D. Huang, "Protein-protein interaction extraction from biomedical literatures based on modified SVM-KNN," in Nat. Lang. Process.Know. Engineer., 2009, pp. 1–7.
[5] H. Kordylewski and D. Graupe, "Applications of the LAMSTAR neural network to medical and engineering diagnosis/fault detection," in Proc7th Artificial Neural Networks in Eng. Conf., St. Louis, MO, 1997.
[6] D. Graupe and H. Kordylewski, "A large memory storage and retrieval neural network for adaptive retrieval and diagnosis," Int. J. Software Eng. Knowledge Eng., vol. 8, no. 1, pp. 115–138, 1998.
[7] Kokol P, Povalej, P., Lenič, M, Štiglic, G.: Building classifier cellular automata. 6th international conference on cellular automata for research and industry, ACRI 2004, Amsterdam, The Netherlands, October 25-27, 2004. (Lecture notes in computer science, 3305). Berlin: Springer, 2004, pp. 823-830
[8] G.Z. Wu, "The application of data mining for medical database", Master Thesis of Department of Biomedical Engineering, Chung Yuan University, Taiwan, Chung Li, 2000.
[9] R. Carvalho, R. Isola, and A. Tripathy, "MediQuery—An automated decision support system," in Proc. 24th Int. Symp. Comput.-Based Med. Syst., Jun. 27–30, 2011, pp. 1–6.
[10] Shucheng Yu, Cong Wang, KuiRen, Wenjing Lou in their paper "Attribute based data sharing with attribute revocation".
[11] C.Y. Hsu, C.S. Lu and S.C. Pei,Image Feature Extraction in Encrypted Domain with Privacy preserving SIFT, IEEE Trans. on Image Processing,21(11): 4593-4607, 2012.
[12] Jun Zhou, Zhenfu Cao, Xiaolei Dong, Xiaodong Lin "PPDM: Privacy-preserving Protocol for Dynamic Medical Text Mining and Image Feature Extraction from Secure Data Aggregation in Cloud-assisted e-Healthcare Systems," IEEE journal of selected topics in signal processing.
[13] Olawuni, Omotayo, Adegoke, Olarinoye ,"Medical image Feature Extraction: A Survey," International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 3 Issue 5 (September 2013)
[14] Peter L. Stanchev, David Green Jr., BoyanDimitrov, "High Level Colour Similarity Retrievl," 28th international conference ICT & P 2003, Varna, Bulgaria.
[15] Jun Zhou, Zhenfu Cao, XiaoleiDond, "Securing M-Healthcare Social Networks: Challenges, Countermeasures and Future Dirctions," IEEE Wireless Communications • August 2013