Personalized Health Information Agent

A.Krishnamoorthy

Abstract— In the future, artificial intelligence machines will replace or enhance human capabilities in many areas. One such a field is Heath care, where in Electronic Health Records (EHR) has been widely used with the help social media to share information about medicines which will cure diseases with others. Artificial intelligence is one rapidly growing technology which helps the researchers to find medical data which are available in the social media and suggest better medicines for the disease queried by the users. Information retrieval is performed by C4.5 algorithm. The system is designed in such a way that the search result will be more optimized with good quality and more efficient to assist the people searching for medical information for particular disease in the internet. The collected information are stored and categorized by using Support Vector Algorithm. Finally, Apriori algorithm is used to make the decision according to the categorized data about the medicines according to the user search.

Index Terms— Artificial Interlligence, Electronic Health Record (EHR); social media; medical data; big data; cluster; analysis.

1 Introduction

Nowadays, electronic health records (EHR) have been widely adopted in hospitals and clinics worldwide. This electronic health records provides significant information about disease details and medical information to the patient posted in the social media (Hanauer et al. 2009; Hanauer et al. 2011; Lin et al. 2011). The report Hanauer et al. (2011) shows the details about used large-scale electronic data, longitudinal electronic health records to research associations in medical diagnosis and consider temporal relations between events to better elucidate patterns of disease progression in the big data. Symptom disease treatment association rule mining is used to identify particular disease information in the big data [Lin et al. (2011)].

Miller 2012b report provides detailed information about health social media sites such as "Daily Strength" and "Patients like Me" websites. Association rule mining and clustering helps to categorize the electronic health records , health social media monitoring and analysis, health text analytics, health ontologies, patient network analysis, and adverse drug side-effect analysis are promising areas of research in the health-related information. Some of them from the following source are listed below,

• Information reliable form online: All the information from online is retrieved from different sources where in the each retrieved information contains different quality. However some sources are not reliable to compare with others. Using new algorithm we can easily make decision comparing to the existing algorithms (C. Sampada, et al, 2004).

• Combined data from various online sources: Search engines provide information based on the user query from various sources of internet; however we need to alter the information to collect required data from it. This process is made more difficult by the usage of different vocabularies and the possible inconsistencies among these information sources.

• Collecting medical data: Once the query is entered into the proposed system means it automatically collects all matching information from the internet about medical data. Identifying and categorizing the medical data is one of the hardest processes. Decision making algorithms helps in this process to identify the required medical data.

• User ability: Users who seek information about medcine for particular disease generally unaware about the retrieved medical information. Most of them are not proficient with information technologies and some of them have limited vision or motor skills caused by aging or illness (Daniel B. Neill).

The above problems are addressed in the proposed system based on providing the functionalities by implementing the algorithms which are proposed. Many of the trusted websites provide medicine information; there are many experts available in the internet to suggest medicine for the users.

Here we are using three algorithms which help to retrieve information from the internet and select matching result of the user query based on the decision making technique and suggest it to the users (C. Sampada, et al, 2004). At the same time we are focused on providing error free proposed system which works efficiently with high speed to suggest medicine details to the users.

2 Overview Of The System Architecture

The above diagram shows the working process of the proposed system. Centralized database is created to store the information about medical data that is retrieved from the internet as per the users query.

2.1 User interface

One of the most important in the proposed system is the UI where users can request and collect information about their medical needs. User interface is designed with simple func-

• *A.Krishnamoorthy is currently pursuing masters' degree program in Computer Science Engineering in Sasurie academy of engineering, Coimbatore. E-mail: akmbtech@gmail.com*

tionalities which help to the users to understand easily. This section simply helps the users to query their problems in the system and provides solutions (S.N. Deepa, B. Aruna Devi, and Nov 2011).

2.2 System understanding

The system need to understand the requirements of the users from their query. So here we used some artificial intelligent techniques which help to identify the user's requirements from their search query and search history. Mostly the system is used to collect information about the medicine for cure from the online electronic health records (Vassilis S Kodogiannis and John N Lygouras (2008)).

2.3 Algorithms

Artificial intelligent health care industry system is using three different algorithms to cluster big data and collecting matching information to suggest the users.

1. C4.5 – To retrieve information from the internet.
2. Support vector machine – Cluster information from Big Data.
3. Apriori Algorithm – It helps to take decision in the collected information.

2.4 Storing information in database

Firstly the artificial intelligent healthcare system collects information from the internet about cure medication details for particular disease. However this information is very high, so the system needs big data base to store and categorize the information. Here the system uses big data technique to store the information and process (Chatterjee, S., & Bandopadhyay, S. 2012).

2.5 Prediction and Description

Prediction and Description plays an important role in the system. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. Description focuses on finding patterns describing the data that can be interpreted by humans [M. Heisler, 2008].

2.6 Trusted data source

Artificial intelligent health care system concentrate on the web search on those trusted data sources which are provided by the healthcare experts and also explicitly consider information reliability in its reasoning process. This system collect information from the trusted data source websites which are identified and the details of the trusted websites are stored in order to filter out the untrusted sources which may have irrelevant data.

Trusted data source are identified based on healthcare experts experience and information is stored in the artificial intelligent health care system database [Kalia Orphanou, Athena Stassopoulou and Elpida Keravnou, March 2014,]. Health care organization experts examine the information that is collected from the internet and suggested to the users. This process will help to avoid some future risks which may occur in the system.

The Artificial intelligence healthcare system is expected to provide assistance to the users or patients in decision making in medicine for instant action. For example if an patient or users suffer some disease in home, artificial intelligent healthcare system suggest information about the medicines which are to be inhaled or whether the user should go to the nearby hospital to undertake take treatment based on its severity[Hu, C., Si, X.-S., & Yang, J.-B. 2010].

At the same time it helps to the users to take decision were they need to go hospital or simply make call to doctor office and just stay at home, it give past health records of the particular disease, artificial intelligent healthcare system provides symptoms of the disease that the users don't have knowledge about it [Chatterjee, S., & Bandopadhyay, S. 2012]. Artificial intelligence healthcare system also provides information for complex health problems with proper treatment information.

3 Hardware and Software Requirements

3.1 Software requirements

Operating System	:	Windows, Linux
Language	:	Java Server Pages
Version	:	JDK 1.7
Back-end	:	MySQL (XAMPP Server)
Server	:	Apache
Tool	:	Net beans, Eclipse

3.2 Hardware requirements

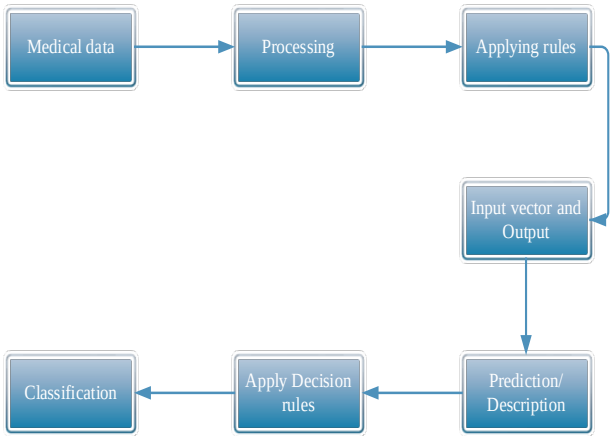
Processor	:	PENTIUM IV
Clock Speed	:	2.7 GHZ
Ram Capacity	:	1 GB
Hard Disk Drive	:	250 GB
Monitor	:	15 VGA Color

4 Rule Generation

The below three steps provides detailed strategies about collecting medical data and process it to provide better solution for the medical relevant queries,

1) Generation of associate thorough set of feasible input vectors for the attention industries, and running of the system to see the output for every input vector,
2) Extraction of information from the set of (input vector, output result) pairs by applying learning or generalization algorithms.
3) Comparison, by associate degree knowledgeable, of the information extracted in step two with the first supply of information (here, the health care industries).

Collecting information from the internet will be difficult for some of the computer users. Because the information in the internet will in different formats and it maybe in huge data size which normal users cannot understand the information will be hidden in the large data set of the medical data. Users will feel very bad to process the whole dataset to get the required information. In that case, the rule generation helps to the system to produce normal computer user understandable information.



The above diagram shows the working of the artificial intelligence Health Care processes. Let us consider an example of the proposed system things like this, user need to input the data to search the matching keywords in the predefined websites to collect information. Once it is completed, then the information collecting process begins and the same is stored. Next step in the process is decision making to classify and filter relevant medical data which are to be suggested to the users. Here proposed system rules will be applied for selecting medical cure information and finally based on the prediction and description part helps to make decision from the decision rules [Mortada, M., Yacout, S., & Lakis, A. 2011]. Completing the above process the artificial intelligent system may classify required information from the users query and suggest it to the users.

The below section provides detailed information's about the algorithms that are used in the artificial intelligent healthcare system.

5 Generating Input Vectors and Outputs

Data mining is the core step, which has resulted in the analyzing of hidden but useful knowledge about medical information from massive electronic health records. Here we propose a non-trivial extraction method which is used to collect information. It can also be defined as "the science of extracting useful information from massive medical social data".

There are two main primary goals which need to be set before cluster and mining the medical data from the big data.

- **Prediction**: it involves some variables or fields in the data set to predict unknown or future values of other variables of interest.
- **Description** focuses on finding patterns describing the data that can be interpreted by humans.

However these techniques help to collect and categorize medical data from the large datasets which are available in the social media and also in the EHR. To provide better treatment information for all the diseases, we need to follow some algorithms to cluster the collected data from the social media and suggest it to the corresponding patients [Rabatel, J., Bringay, S., & Poncelet, P. 2011]. Data mining have a collection of algorithmic ways to extract informative patterns from raw data – Data mining is purely data-driven; this feature is important in health care,

1) $y = f(x)$
2) We have seen x (set of independent variables) and observed y (dependent variable); data mining tells us something about the nature of f
   - x = symptoms or test results, y = diseases;
   - x = treatments, y = symptom
3) It tells us "how"
   - How is x related to y? What function describes their relationship?
4) $f(x, y) = \text{score}$, or $f(x \mid y) = Pr(x \mid y)$
5) Data mining does not (directly) explain to us "why" – Why does x cause y?
6) It helps doctors/physicians (domain experts) figure that (causation) out
7) 'Descriptive/predictive model' vs. 'Causal model'

6 Building the Decision Tree

We propose, C4.5 to collect information that is related to the Healthcare medical data in big data, because C4.5 provides accurate results for any problems. By using this we can collect user inputs and retrieve the relevant medical information available in the big data and we can provide better solutions. Firstly this C4.5 is been developed to classify and to form the decision tree. Classifier is known as tool which helps in the data mining process to take large amount of data representing things we want to classify and attempts to predict which class the new data [Si, X. S., Hu, C. H., Yang, J. B., & Zhang, Q. 2011].

A decision tree is built from the input vectors and the associated outputs, using C4.5, a reference algorithm in machine learning. Pruning must be disabled, to ensure 0% error in the tree [C. Sampada, et al, 2004]. Factorization rules are applied to reduce the size of the tree:

1) If all the children of a given node include the same element of a recommendation (e.g. a recommended drug treatment), this information can be included in the node and removed from its children,
2) If a particular variable in medical data takes several values leading to the same recommendations, the largest set of such values can be grouped together as "<other>".

The important steps in the C4.5 decision tree,

- First, C4.5 collects information those are related to our requirements and create decision tree corresponding to the data.
- Second, C4.5 uses a single-pass pruning process to mitigate over-fitting the new data. And finally it produces results in many improved formats as developer requirement.
- Third, it has the ability to work on both continuous and discrete data.
- Finally, incomplete data is deal with in its own ways.

Using this C4.5 technique we can easily collect data that is related to our healthcare system and processing this information is very easy.

Pseudo code

The general algorithm for building decision trees is:

1) Check for base cases
2) For each attribute a
   2.1) Find the normalized information gain ratio from splitting on a
3) Let a_best be the attribute with the highest normalized information gain
4) Create a decision node that splits on a_best
5) Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node

7 Comparing the Decision Tree with Clinical Guidelines

Once the C4.5 decision tree process is completed then it helps to collect the medical data from the patient's and expert's data which is stored in the big data. To make decision about suggesting cure medicine information from collected data, the proposed system can use Apriori algorithm which provides better results comparing with other techniques. The Aprior algorithm provides accurate results or solution for any kind of problem as per the researcher requirements. The proposed system algorithm uses keyword which helps to retrieve information from the database and cluster those information's as per the keywords [Z. H. Deng and S. L. Lv.]. The algorithm important steps are given below,

Join Step: Ck is generated by joining Lk-1 with itself

Prune Step: Any (k-1) - item set that is not frequent cannot be a

- Subset of a frequent k - item set.
- Ck: Candidate item set of size k
- Lk: frequent item set of size k
- L1 = frequent items
- for (k = 1; k < L; k++) do begin
  - Ck+1 = candidates generated from Lk
  - for each transaction t in database do
    - increment the count of all candidates in
    - Ck+1 that are contained in t
- Lk+1= candidates in Ck+1 with min support
- end
- return Lk U Lk

8 The Exact Output from Clustering

The proposed system sample screens which shows the information about the process of the artificial intelligence in health advice system.

However to verify the results of the system we need to add some important points in the big data to cluster and process other techniques. Firstly, user of the proposed system needs to enter the cause information in the system.

This information helps to the admin to tract the user's requirements and continuously searched details in the healthcare system.

At the same time users of the proposed system need to enter symptoms to check the information and cure medicine details for particular symptom in the system.

The system collect diagnosis details from the different internet sources and save it in the local server to provide diagnosis details to the users. If any diagnosis details are not available in the system data base then admin need to collect and need to enter manually in the proposed system.

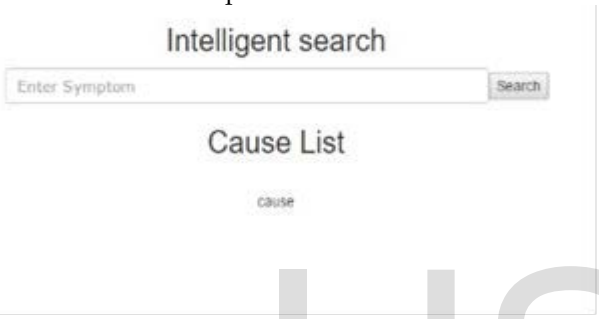
Every time admin need to check and differentiate the electronic health record information in the proposed system, the below figure show the process of differentiating the EHR in the system.

Finally users get search box in the artificial intelligent healthcare advice system to collect information about particular cure medicine details through the system. By entering the

required information's as a query in the search box, the system will automatically collect.

Once user completed the above process then the system collect information that are relevant to the users query and cluster information from the medical data sets. Using the support vector machine technique and apriori algorithm the proposed system analyzes the details in the datasets and suggest to the users.

At the same time it lists the medical data information with the cause details and symptom of particular disease with diagnose and source of the particular information.



8.1 Support Vector Machines (SVM)

Support Vector Machines (SVM) has many advantages, and some of the advantages are given in [24] they are: support vector machines produce an accurate classification results on the theoretical basis, even when input data are non-linearly separable. However, the predicted result does not rely on the quality of human expertise judgment for choice of the linearization function for nonlinear input data.

Some of the disadvantages of SVM as a non-parametric technique introduced by [24] are its absence for transparency for results. One of the main disadvantage of the support vector machine is discussed by [25] is that which lies in the choice of the kernel. It has to be set correctly to achieve an accurate result for any given task or problem. Kernel choice that produces accurate results for task A "may produce poor results for task B".

Support vector machine technique helps to create training sets and prediction models which plays important role in identifying the exact matches.

- input/output sets X , Y
- training set (x1, y1),...,(xm, ym)
- "generalization": given a previously seen $x \in X$, find a suitable $y \in Y$.
- i.e., want to learn a classifier: y=f(x, α), where α are the parameters of the function.
- For example, if we are choosing our model from the set of hyperplanes in Rn, then we have:

$$F(x, \{w, b\}) = \text{sign}(w \cdot x+b).$$

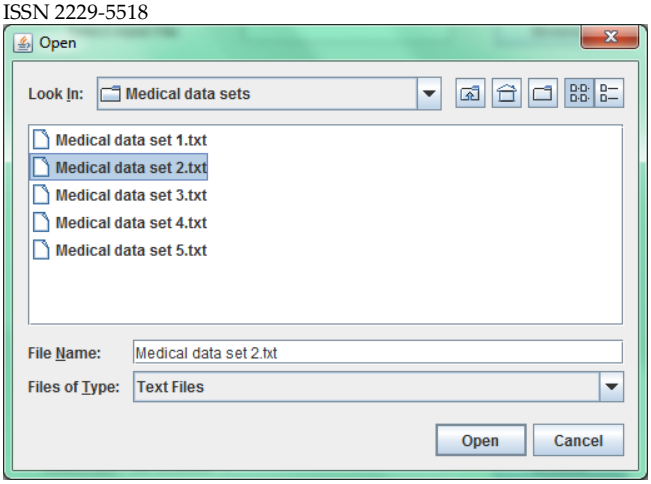
Using this technique the proposed system can easily cluster required information form the collected big data using the training data sets. However, the proposed system takes many processes to cross validation and grid search to find the matching keywords in the database. Collecting data sets from the internet resources,

```
Public class GetsSets {
    Private static GetsSets gs;
    Public static GetsSets getInstance(){
    if(gs == null)
            gs = new GetsSets();
    return gs;
    }
    Private GetsSets(){
    }
}
```

8.2 The design and working principles of apriori algorithms

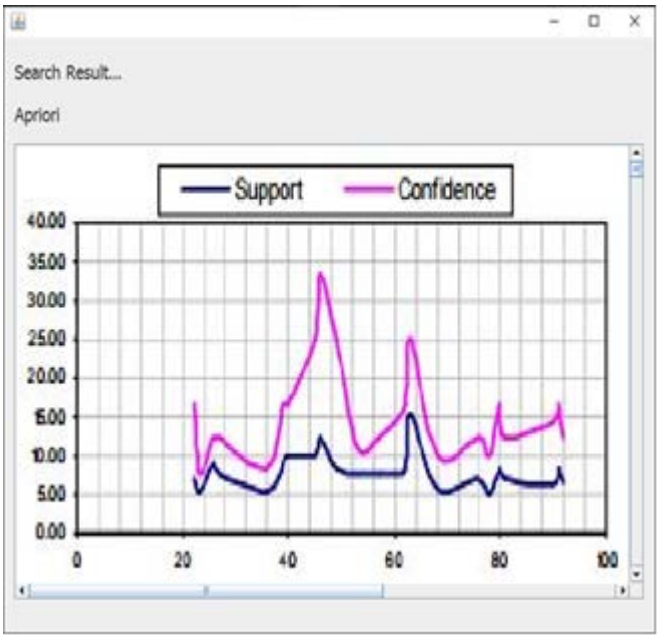
Firstly, the proposed system prcesses the data set of the medical information to cluster the useful required information. After selecting the data set file the system process different steps in it.

Selecting medical data set file to process,



Processing medical data in the proposed system to cluster the information.

Final results of the apriori algorithm.



The above screens shows the overall process of the appriori algorithm and working principles to cluster usefull medical information form the collected medical data of the system. once the above process completed then the system suggest sollution to the users of the proposed system.

Private boolean checkWordPresence (List<String> tokenizedReviewList, String inputWord){

```
Boolean flag = true;

int count = 0;

for(String review : tokenizedReviewList){
for(String word : Arrays.asList(review.split(" "))){
if(word.equalsIgnoreCase(inputWord)){
count++;
if(count < 11)
flag = false;
return flag;
}
```

```
private Set<String> getWordSet(List<String> tokenizedReviewList){
Set<String> uniqueWordSet = new LinkedHashSet<>()
tokenizedReviewList.forEach(
(review) -> {
Arrays.asList(review.split(" ")).forEach(
(word) -> {
word = word.trim();
if(!(word.equals("") || word.equals("\n") || word.equals(" "))){
//check it is integer alone
if(!word.matches("^(?:(?:\\-{1})?\\d+(?:\\.{1}\\d+)?)$")){
//checking string length more than 2
if(checkWordPresence(tokenizedReviewList,word))
uniqueWordSet.add(word);
return uniqueWordSet;
}
```

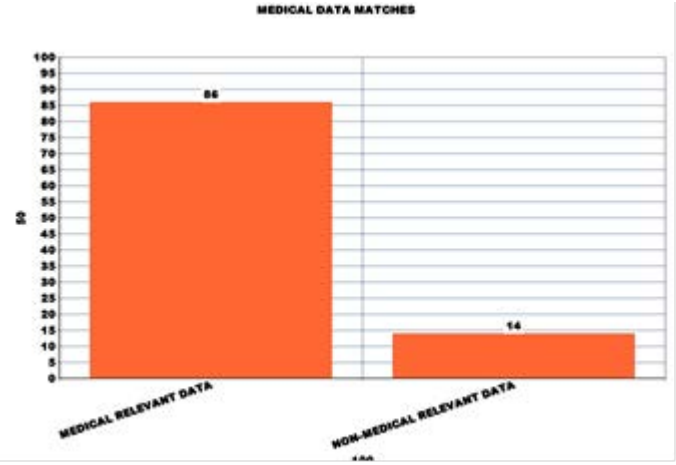
9 RESULT VERIFICATION

The sections help to the researcher to verify the results of the proposed system. Whether it produces expected results of the researchers from the collected medical data and suggest to the users correctly or not. This process helps to the medical experts to verify the medical data matches and suggested medicines for particular disease to users are matching with their suggested cure medicine. The proposed system collects two type of data or information from the internet, one is medical relevant information and other one is non-medical relevant information but somehow they related with the medical information (S. A. Becker, 2004).

Users need to provide exact information about the disease symptoms which helps to the artificial intelligent system to serve users by providing better medical cure information. Disease is posted up on the screen. A "minus" of this artificial intelligent system (and usually of any other Expert System) is that only the symptoms of disease need to put in the knowledge base by the programmer are available. It doesn't think and doesn't learn by itself; but the knowledge

base can be updated anytime with new symptoms and new diseases.

This chart represents the accuracy of the data. The information provided by the proposed system is examined by the experts in the medical field. This process helps to improve the quality of service.



10 Conclusion and Future Work

The implementation of the artificial intelligent healthcare system dynamically collect information from online knowledge resources as well it collect personalized information of disease cure medicine from the Electronic Health Records (HER) from internet. The importance of this research is by no means to restore medical expert to perform diagnosis or critical treatment decision-making, given the complexity, liability and law requirement in medical practice domain. This research is mainly concentrated in helping the users who all are searching cure information in internet. There is massive health related information is available online given by the patients personal background information including their health problems which helps to others to understand easily. At the same time information about patients experience in taking treatment for particular disease and treatment procedure including cure medicine information. These data helps to the users to understand the health problem and taking treatment for the corresponding disease with cure medicine information.

On the other hand, it is possible that the symptoms already present are not 100% right, because different experts have different opinions and there are a lot of anomalies in medicine. This artificial intelligent healthcare system is improved to provide various services in the medical field. The important technique that helps to improve this artificial intelligent healthcare system to assist users or patients is decision making algorithms. However In this situation the proposed system would establish the degree that the diagnosis is close to the reality.

Acknowledgment

I would like to extend my sincere thanks to all people mentioned below who have helped me in carrying out this research. I record my deep sense of indebtedness and whole hearted gratitude to head of the department, Mrs. N. Poongothai M.E., Assistant Professor and Head, Department of Computer Science and Engineering, Sasurie Academy of Engineering, Coimbatore, for her active involvement, encouragement, caretaking and valuable suggestions in all the way and also shaping the research work and this report. I am highly indebted to Mr. S. Sakthivel M.E, Assistant Professor, Department of Computer Science and Engineering for his support, guidance and constant supervision during this work.

Finally, I thank all my classmates for their kind cooperation and encouragement.

References

[1] Hanauer, D. A., Rhodes, D. R., and Chinnaiyan, A. M. 2009. "Exploring Clinical Associations Using '-Omics' Based Enrichment Analyses," PLoS ONE (4:4): e5203.

[2] Hanauer, D. A., Zheng, K., Ramakrishnan, N., and Keller, B. J. 2011. "Opportunities and Challenges in Association and Episode Discovery from Electronic Health Records," IEEE Intelligent Systems (26:5), pp. 83-87.

[3] Lin, Y., Brown, R. A., Yang, H. J., Li, S., Lu, H., and Chen, H. 2011. "Data Mining Large-Scale Electronic Health Records for Clinical Support," IEEE Intelligent Systems (26:5), pp. 87-90

[4] Machine learning and Data Mining - Association Analysis with Python FRIDAY, JANUARY 11, 2013 http://aimotion.blogspot.in/2013/01/machine-learning-and-data-mining.html

[5] Miller, K. 2012a. "Big Data Analytics in Biomedical Research," Biomedical Computation Review (available at http:// biomedicalcomputationreview.org/content/big-data-analyticsbiomedical-research; accessed August 2, 2012).

[6] Miller, K. 2012b. "Leveraging Social Media for Biomedical Research: How Social Media Sites Are Rapidly Doing Unique Research on Large Cohorts," Biomedical Computation Review (available at http://biomedicalcomputationreview.org/content/ leveraging-social-media-biomedical-research; accessed August 2, 2012).

[7] Gelfand, A. 2011/2012. "Privacy and Biomedical Research: Building a Trust Infrastructure—An Exploration of Data-Driven and Process-Driven Approaches to Data Privacy," Biomedical Computation Review, Winter, pp. 23-28 (available at http://biomedicalcomputationreview.org/content/privacy-andbiomedical-research-building-trust-infrastructure, accessed August 2, 2012).

[8] Charles Weddle, Graduate Student, Florida State University "Artificial Intelligence and Computer Games", unpublished.

[9] C. Sampada,, et al, 2004, "Adaptive Neuro-Fuzzy Intrusion Detection Systems", Proceedings: International Conference on Information Technology: Coding and Computing (ITCC'04).

[10] Daniel B. Neill, "Using Artificial Intelligence to Improve Hospital InpatientCare".

[11] S.N. Deepa, B. Aruna Devi, (Nov 2011), "A survey on artificial intelligence approaches for medical image classification", Indian Journal of Science and Technology, Vol. 4 No. 11.

[12] Vassilis S Kodogiannis and John N Lygouras (2008) Neuro-fuzzy classification system for wireless capsule endoscopic images.

[13] S. A. Becker, 2004, "A study of web usability for older adults seeking online health resources," ACM Transactions on Computerhuman Interaction, vol. 11, no. 4, pp. 387–406.

[14] Chatterjee, S., & Bandopadhyay, S. (2012). Reliability estimation using a genetic algorithm-based artificial neural network: An application to a load-haul-dump machine. 39(12), 10943–10995.

[15] M. Heisler, 2008, "Actively engaging patients in treatment decision making and monitoring as a strategy to improve hypertension outcomes in diabetes mellitus." Circulation, vol. 117, no. 1355-1357.

[16] Kalia Orphanou, Athena Stassopoulou and Elpida Keravnou, March 2014, "Temporal abstraction and temporal Bayesian networks in clinical domains: A survey," Artificial Intelligence in Medicine, vol. 60, issue 3, pp. 133-149.

[17] Hu, C., Si, X.-S., & Yang, J.-B. (2010). System reliability prediction model based onevidential reasoning algorithm with nonlinear optimization. Expert Systems with Applications, 37(3), 2550–2562.

[18] Wang, C., & Lin, T. (2011). Improved particle swarm optimization to minimize periodic preventive maintenance cost for series-parallel systems. Expert Systems with Applications, 38(7), 8963–8969.

[19] Mortada, M., Yacout, S., & Lakis, A. (2011). Diagnosis of rotor bearings using logical analysis of data. Journal of Quality in Maintenance Engineering, 17(4), 371–397.

[20] Rabatel, J., Bringay, S., & Poncelet, P. (2011). Anomaly detection in monitoring sensor data for preventive maintenance. Expert Systems with Applications, 38(6), 7003–7015.

[21] Si, X. S., Hu, C. H., Yang, J. B., & Zhang, Q. (2011). On the dynamic evidential reasoning algorithm for fault prediction. Expert Systems with Applications, 38(5), 5061–5080.

[22] Z. H. Deng and S. L. Lv. 2014, Fast mining frequent itemsets using Nodesets.[2]. Expert Systems with Applications, 41(10): 4505–4512.

[23] Ramezani, Reza, Mohamad Saraee, and Mohammad Ali Nematbakhsh; 2014, MRAR: Mining Multi-Relation Association Rules, Journal of Computing and Security, 1, no. 2.

[24] L. Auria and R. A. Moro, "Support vector machines (SVM) as a technique for solvency analysis," Discussion Papers of DIW Berlin, no. 811, 2008.

[25] S. Martin. Disadvantages of support vector machines (SVM). [Online]. Available: http://www.svms.org/disadvantages.html.