AI ALGORITHM FOR HEALTH RISK PREDICTION

Shubhangi Ghodake
Department of Computer Engineering
JSPM Narhe Technical Campus, Narhe, Pune
ghodkeshubhangi12@gmail.com

Dr. Nihar Ranjan
Department of Computer Engineering
JSPM Narhe Technical Campus, Narhe, Pune
nihar.pune@gmail.com

Abstract-In general public, individuals have large concern for their own health. Personalized health is step by step increasing. The lack of experienced specialists as well doctors, many of the healthcare organizations cannot meet the medical demand of patient/public. Patient need more accurate and instant result. Hence, many data mining applications are develop to gives people more alternative healthcare service. It is a qualitative solution for the resolving conflict between less available medical resources and increasing medical demands. In Our System using data mining methods to discover the co-relation between the regular physical symptoms and the possible health risk given by the user or patients. The Main concept to identify medical diseases according to given symptoms ,daily Routine when User search the hospital then recommend the nearest hospital of their current location. The application provides a user-friendly UI for patient and doctors. Patients knows their symptoms which observed in body while doctors can get a number of patients with possible risk. Patients can book appointment of doctor. A feedback mechanism could save human resources and enhance performance of system automatically. The doctor fix prediction result given by system through a user interface, which will collect doctors' advice as new training data. Thus, our application increase the execution and performance of prediction model automatically.

Keyword: Prediction, Keyword Searching, Baseline, XGboost, Query Suggestion.

1. INTRODUCTION

Many healthcare are busy in assisting people with best-effort healthcare service. Now a days, individuals have more concern for their own health. They want higher valuable, qualitatively and personalized healthcare solutions. However, with less of number of skilled doctors and physicians, most healthcare cannot able to complete the requirement and need of patients. How to provide higher quality healthcare to many patient with small number of human resources becomes a main problem. The healthcare environment is normally identify as being 'information rich' yet 'knowledge poor' [1]. Hospital information systems mostly generate large amount of data which takes in the form of text numbers etc. There is a most of hidden information in given data not touched. Predictive analytic and data mining focus on to identify patterns and rules by applying data analysis techniques on a big set of data for detailed and predictive purposes. Data mining is useful for processing and analyzing large datasets from hospital information system and finding relations among data features. Data mining applications in healthcare can be collect as the estimation into large n of categories [15]. It takes minimum no of researchers to analyze data from hospital data. The Main purpose to identify medical diseases according to given symptoms and daily Routine when User search the hospital then recommends the nearest hospital of their current location. The application provides a user-friendly UI for patient and doctors. Patients can know their symptoms which observed in body while doctors can get a number of patients with possible risk. A feedback mechanism could save time, human resource & enhance execution & performance of application automatically.

1.1 Motivation

1. Previous medical examiner only used mostly basic symptoms of particular diseases but our system examiner examines on the word count, questionnaire and diagnostic data.
2. A feedback mechanism which save time, human resources & enhance performance of system automatically. The doctor fix prediction result through an interface, which will usually gather and collect doctors' input as new created training data. An extra training process will be daily apply and trigger on system using these given data. Thus, our system could enhance, improve the performance of prediction model automatically.
3. When the user visits doctor & hospital physically, then user's personal record is collect, saved and then that record is used in the examiner data set. It takes lot of time for this process. Our system reduces waiting time.
4. Present system work on only single hospital, our system work on the data-set of different hospital.

1.2 Objective

1. To determine medical diseases according to daily routine & given observed symptoms.
2. When user search the hospital by keyword, the hospital which is nearest to their current location is recommend.
3. Develop Feedback mechanism to increase and enhance performance of system automatically. To give option of booking of doctor appointment.

2. RELATED WORK

[1] **"Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" Author,-Srinivas K, Rani B K, Govrdhan A.** The social insurance condition is generally seen as being 'data rich' yet 'information poor'. There is bounty of data open inside social protection structures. There is a nonappearance of fruitful examination gadgets to find disguised associations and examples in data. Data disclosure and data mining have create different applications in business and consistent region. Critical taking in can be found from use of data mining systems in human administrations structure. In this examination, rapidly investigate the potential usage of portrayal based data mining techniques, for instance, rule based, decision tree, artless bayes and phony neural framework to gigantic volume of social protection data. The human administrations industry accumulates enormous proportions of social protection data which, tragically, are not "mined" to discover hid information. This is a development of blameless Bayes to questionable probabilities that goes for passing on incredible requests in like manner while overseeing close to nothing or divided instructive files. Disclosure of hid precedents and associations every now and again goes unexploited. It enables vital learning, for instance plans, associations between therapeutic parts related to coronary disease, to be set up.[2]

[2] **"Grand challenges in clinical decision support" Author- Sittig D, Wright A, Osheroff J, et al.** There is a crushing prerequisite for huge gauge, effective strategies for organizing, making, showing, executing, surveying, and keeping up a wide scope of clinical decision help capacities for clinicians, patients and purchasers. Using an iterative, understanding building process we perceived a rank-asked for once-over of the primary 10 thousand inconvenience in clinical decision help. This summary was made to instruct and move authorities, specialists, funders, and approach makers. The once-over of troubles orchestrated by essentialness that they be comprehended whether patients and affiliations are to begin understanding the fullest points of interest possible of these systems involves: upgrade the human– PC interface; dissipate best practices in CDS structure, enhancement, and utilization; shorten calm measurement information; sort out and channel recommendations to the customer; make a building for sharing executable CDS modules and organizations; merge proposition for patients with co-morbidities; arrange CDS content progression and execution; settle on web accessible clinical decision enable chronicles; to use free substance information to drive clinical decision assist; mine significant clinical databases with making new CDS. Conspicuous confirmation of answers for these challenges is essential if clinical decision help is to achieve its potential and upgrade the quality, security and capability of therapeutic administrations.[3]

[3] **"Using Electronic Health Records for Surgical Quality Improvement in the Era of Big Data" Author-Anderson J E, Chang D C.** Many Improvement in the Era of Big Data" Author-Chang D C.,Anderson J E}Various social protection workplaces approve security on their electronic prosperity records through a healing segment: some staff apparently have for all intents and purposes boundless access to records, yet there is strict ex post facto audit process for wrong gets to, i.e., gets to that dismiss the workplace's security and insurance approaches. This system is inefficient, as each suspicious access must be explored by a security ace, and is totally audit, as it occurs after damage may have been obtained. Past undertakings at such a system have successfully associated overseen learning models to this end, for instance, SVMs and key backslide. While giving points of interest over manual reviewing, these strategies dismiss the identity of the customers and patients related with a record get to. Along these lines, they couldn't manhandle the way that a patient whose record was as of late drawn in with an encroachment has an extended peril of being related with a future encroachment. Prodded by this, in this paper, propose a common filtering moved approach to manage predicting uncouth gets to. Our answer arranges both express and inert features for staff and patients, the last going about as a tweaked "remarkable imprint" in light of chronicled get to structures. The proposed method, when associated with certifiable EHR get to data from two tertiary specialist's offices and a record get to dataset from Amazon, exhibits not simply basically expanded

execution appeared differently in relation to existing procedures, yet what's more gives bits of information in regards to what demonstrates a wrong access.[4]

[4] **"Data Mining Techniques into Telemedicine Systems" Author-Gheorghe M, Petre R** giving consideration benefits through telemedicine has turned into an imperative piece of the therapeutic improvement process, because of the most recent advancement in the in-development and PC innovations. Then, information mining, a dynamic and quick growing area, has enhanced numerous fields of human life by offering the likelihood of anticipating future patterns and assisting with basic leadership, in view of the examples and patterns distinguished. The decent variety of information and large number of information mining methods give different applications to information mining, incorporating into the social insurance association. Coordinating information mining strategies into telemedicine frameworks would help enhance the productivity and viability of the social insurance associations movement, adding to advancement and refinement of medicinal services administrations offered as a component of the restorative improvement process.[5]

[5] **"Query recommendation using query logs in search engines" Author-R. Baeza-Yates, C. Hurtado, and M. Mendoza** In this paper propose a technique that, given an inquiry submitted to a web list, prescribes a once-over of related request. The related inquiries are arranged in advance issued request, and can be issued by customer to the web record to tune or redirect the interest method. Procedure proposed relies upon an inquiry clustering process in which social occasions of semantically similar inquiries and issues are perceived. The gathering method uses the substance of chronicled tendencies of customers selected in the request log of web list. The system finds the related inquiries, just as positions them as shown by congruity standard. Finally, we show up with examinations over the inquiry log of a web file the practicality of the strategy.[6]

[6] **"Data Mining Applications In Healthcare Sector: A Study" Author -M. Durairaj, V. In** this paper, system have focused to examine a grouping of techniques, approaches and different instruments and its impact on the social protection part. The goal of data mining application is to turn that data are substances, numbers, or substance which can be set up by a PC into learning or information. The guideline inspiration driving data mining application in human administrations systems is to develop a mechanized gadget for perceiving and scattering essential social protection information. This paper intends to make a dirty examination report of no of sorts of data mining applications in the human administrations division and to diminish the multifaceted idea of the examination of the social protection data trades. In like manner shows a comparative examination of various data mining applications, systems and unmistakable approaches associated for expelling gaining from database made in therapeutic industry. Finally, current data mining methodologies with data mining estimations and its application mechanical assemblies which are logically beneficial for social protection organizations are analyzed in detail.[7]

[7] **"Detecting Inappropriate Access to Electronic Health Records Using Collaborative Filtering" Author-Aditya Krishna Menon,** No of therapeutic administrations workplaces approve security on their electronic prosperity records through a remedial framework: some staff apparently have for all intents and purposes boundless access to the records, yet there is strict ex post facto survey process for classless gets to, i.e., gets to that dismiss the workplace's security and assurance approaches. This system is inefficient, as each suspicious access must be surveyed by a security ace, and is basically audit, as it occurs after damage may have been achieved. In this way, they couldn't abuse the way that a patient whose record was as of late connected with an encroachment has an extended peril of being related with a future encroachment. Stirred by this, in this paper, propose an agreeable filtering breathed life into approach to manage predicting inappropriate gets to. Answer consolidates both unequivocal and latent features for staff and patients, the last going about as an altered "special finger impression" in perspective of obvious access plans. The proposed procedure, when associated with certified EHR get to data from two tertiary specialist's offices and a report get to dataset from Amazon.[8]

[8] **"Text data mining of aged care accreditation reports to identify risk factors in medication management in Australian residential aged care homes" Author-Tao Jiang & Siyu Qian,** This examination intended to recognize peril factors in medication the officials in Australian private developed thought (RAC) homes. Only 18 out of 3,607 RAC homes failed developed thought accreditation standard in medication the officials between seventh March 2011 and 25th March 2015. Content data mining procedures were used to explore purposes behind dissatisfaction. This provoked the unmistakable verification of 21 chance pointers for RAC home to bomb in medicine organization. These pointers were moreover collected into ten subjects. They are in explicit

medication the board, tranquilize examination, asking for, directing, limit, stock and exchange, association, event report, watching, staff and occupant satisfaction. The best three danger factors are: "insufficient watching method" (18 homes), "disobedience with master standards and guidelines" (15 homes), and "occupant dissatisfaction with for the most part talking medicine organization" (10 homes).[9]

[9] **"Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods" Author-Tim W. Nattkempera,, Bert Arnrich** The k-infers clustering and self-dealing with maps (SOM) are associated with analyze banner structure similarly as representation. They use k-nearest neighbor classifiers (k-nn), reinforce vector machines (SVM) and decision trees (DT) to portray features using PC helped discovering (CAD) approach.[10]

[10] **"Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses" Author-Song J H, Venkatesh S S, Conant E A,** Chest sickness is a champion among the most generally perceived malignancies in women. Sonography is at present customarily used in blend with various moralities for imaging chests. Notwithstanding the truth that ultrasound can examine direct bruises in the chest with an accuracy of 96 rate – 100 rate, its usage for unequivocal partition between solid thoughtful and unsafe masses has ended up being progressively troublesome. Despite amazing undertakings toward improving imaging systems, including sonography, last attestation of whether solid chest damage is undermining or kind is up 'til now made by biopsy.[11]

[11] There are large no of healthcare applications based on data mining. Logistic regression models are generally used to compare hospital account based on risk-adjusted death with thirty days of non-cardiac surgery [13]. The purpose of an MDDS is to raised, not replace, the natural ability of human diagnosticians in the complex process of medical diagnosis [12].The procedure & idea of developing a new data mining technique & software to help competent solutions for medical data analysis has been described. It Suggest a hybrid tool and techniques that incorporates RST and ANN to make proficient data analysis and indicative predictions [14]. Data mining applications in healthcare can be collect as the estimation into large n of categories. Data mining applications can develop to estimate the success of medical therapy. Big amounts of data are a main resource to be processed & examined for knowledge extraction which allow to support for decision making and cost reduce [15,16]. Different data mining methods that have been utilized for breast cancer detection and prognosis Decision tree is found to be the god predictor with 93.62 Accuracy on benchmark dataset and also on SEER data set [17]. Data mining analysis implement on given such data may not only reveal new information on the detected diseases, it also more optimize settings of the applied diagnostic tests as well. [18]

3. SYSTEM ARCHITECTURE

In our general public, individuals have more concern for their own health. Personalized health is step by step rising. The lack of experienced specialists and doctors, many healthcare organizations cannot meet the medical demand of people. Patient want more correct and instant result. Hence, many data mining applications are develop to gives people more alternative healthcare service. It is a qualitative solution for the resolving conflict between less available medical resources and increasing medical demands. In Our System using data mining methods to discover the co-relation between the regular physical symptoms and the possible health risk given by the user or patients. The Main Concept to determine medical diseases according to given symptoms ,daily Routine when User search the hospital then recommend the nearest hospital of their current location. The Main Concept to determine medical diseases according to given symptoms ,daily Routine when User search the hospital then recommend the nearest hospital of their current location. The application provides a user-friendly UI for examines and doctors. Patients can know their symptoms which occurred in body while doctors can get a set of patients with possible risk. Patients can book appointment of doctor.A feedback mechanism could enhance performance and save human resources of application automatically. The doctor fix system prediction and classification result through an UI, which will gather doctors' advice as new training data. Thus, our system could increase the execution and performance of prediction model automatically. It give option of booking doctor appointment.

3.1 Advantage of proposed

- Improve human-computer interactions.
- User and nearest hospital location is detected.
- Recommended the doctor and hospital to patient according to diseases Predicted.
- Improved performance by using feedback.
- Scalable, fast, low cost Prediction system.
- Comparable and good quality to experts.

- Online Doctor Appointment system provided.

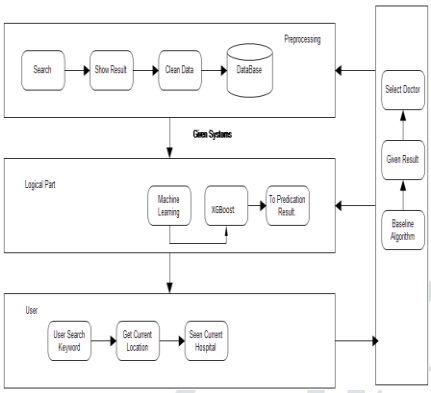

Fig 1: System Overview

- Users are first registration and use first lo-gin with OTP (one time password) and next time only lo-gin without OTP (one time password).
- After lo-gin users have 2 options first one is to give symptoms to the application, disease will predict and second one Search hospital using keyword and nearest hospital location is recommended.
- In First XgBoost and decision algorithm is used for disease prediction.
- In second one Partitions Algorithms, baseline Algorithms which User Search With keyword. Admin Lo-gin and after admin add hospital and Check user Details. Doctor First Register with select hospital and lo-gin after lo-gin Doctor show the Details of appointment and feedback is provided for predicted result.

3.2 Mathematical Model

$$\sum_{n=1}^{\infty} |\mathbf{O(E)} - \mathbf{O1(E1)}| \quad (1)$$

Here in equation 1, |O(E)} is Training Data Set which train by our System. And |O1(E1)| is Input Data Set given by User.

We define a mapping O from E to C = {0, 1}, where C represents whether a examinees Show to result According to give Input. Our task is to find an appropriate model *f : E1O()*. f can predict whether examinees result.

S= { }

Identify the inputs as symptoms and search hospital.

F= {f1, f2, f3 ....., fn — F as set of functions to execute similar diseases and medicine}

I= {i1, i2, i3 —I sets of inputs to the symptoms}

O= {o1, o2, o3. —O set of outputs from the function sets}

S= {I, F, O} I = {symptoms and search hospital, Search By doctor name, search by specialization}

O = {predicted the diseases and medicine, Locate the current location}

F = {Keyword Search, Text Processing, Prediction techniques, Baseline Algorithms, XGBoost}

3.3 ALGORITHMS

1. XGBoost

XGBoost is optimize distributed gradient boosting library designed to be most efficient, adjustable and portable. The library is mainly focused on computational speed and model execution, performance. XGBoost gives a parallel tree boosting (also known as GBM, GBDT) that solves most of data science problems in a quick and correct way. Gradient boosting is a machine learning technique which is used for regression and classification problems, which generate a prediction model in the form of collection of weak prediction models, generally decision trees. It develop and build the model in a stage wise fashion like other boosting algorithm methods are developing, and it generalizes them by granting allowing optimization of arbitrary differential loss function. In boosting, no of trees are built sequentially, one by one such that each subsequent tree has aims to reduce the errors of the previous tree. Each tree learns from its previous and updates the errors. Hence, the tree that grows and develop next in the sequence it will learn from an updated version of the residuals.

Like any other boosting methods, gradient boosting is combines weak "learners" into the single strong learner in an iterative fashion. It is simplest to explain in the least-squares regression setting, where the goal is to "teach" "learn" a model *F* to predict values of the form

y' = F(x)y' = F(x) (2)

By minimizing the mean squared error

$$In \sum_i (y'i - yi)^2 \;\; In \sum_i (y'i - yi)^2 \quad (3)$$

Commented [SBG1]:

where i At each stage m, $1 \leq m \leq M$ of gradient boosting, it may be assumed that there is present some imperfect model Fm (at the outset, a very weak model that just predict mean y in training set could be used). The gradient boosting algorithm improves on Fm by constructing new model that Adds an estimator h to provide the better model:

$$Fm+1(x) = Fm(x) + h(x) \quad (4)$$

To find h, the gradient boosting solution initiate with the observation that a perfect h would imply.

$$Fm+1(x) = Fm(x) + h(x) = y \quad (5)$$

Or, equivalently,

$$h(x) = y - Fm(x) \quad (6)$$

2. Partition based algorithm

Algorithm BA can be slow for different reasons. Initially, at each iteration, only one node is processed; thus, the active ink drops slowly and termination conditions are met after too much iteration. Second, given large number of iterations, the overhead of maintaining queue is significant. Then finally, all nodes distribute their active ink to all their neighbors, even if some of them only receive a small amount of ink. To improve performance of BA, in this section, we propose partition-based algorithm which divides the keyword queries and the documents in the KD-graph G into groups.

- Implement algorithm and test it.
- Instrument algorithm so that it counts number of comparisons of array elements. (Dont count any comparisons between array indices.) Test it to see if the counts make sense or not.
- For given values of n from 500 to 10000, run number of experiments on randomly-ordered arrays of size n and find average number of comparisons for those experiments.
- Graph average number of comparisons as function of n. Repeat given above items 14, using an alternative pivot selection method.

3. Baseline algorithm

Book Mark Coloring Algorithm is previous Algorithm used to calculate distance and time then Baseline Algorithms is used because BA is initiate with one unit of active ink injected into node kq, BA processes all nodes in the graph in descending order of their active ink. Different from typical personalized PageRank problems where graph is homogeneous, our KD graph Gq has the two types of nodes: keyword query nodes and Search nodes. As opposed to BCA, BA only ranks keyword query nodes; keyword query node retains the portion of its active ink and distributes 1- alpha portion to its neighbor nodes based on its outgoing adjusted edge weights, while the document node distributes all its active ink to its neighbor nodes.

- Implement algorithm and test it.
- To find RWR based top-m query recommend.
- Start from one unit of active ink injected into node Kq and order in descending order.
- Find weight of each edge e is adjusted based on q.
- The algorithm returns top-m candidate suggestions other than kq in C as the result.

4. SYSTEM ANALYSIS AND RESULT

In Our System divide the data set into training set and testing set by 2 to 1. Then Our System apply the two algorithms mentioned above in risk-prediction task of three symptoms. These three symptoms system ask the some Question about the symptoms when System ask question and User are given the answer. In Our System User Search with keyword like specialization doctor name and hospital name and Get nearest hospital Result according to current Location.

Comparison between Algorithms:

Sr.No	ALGORITHMS	Time(In Milliseconds)
01	Baseline(Same Area)	20000
02	Baseline(Different Area)	36036

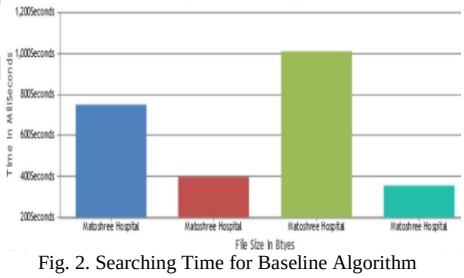

Fig. 2. Searching Time for Baseline Algorithm

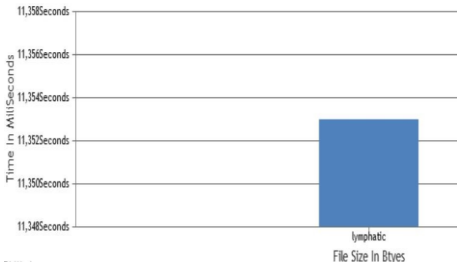

Fig. 3. Searching Time for XGBoost Algorithm

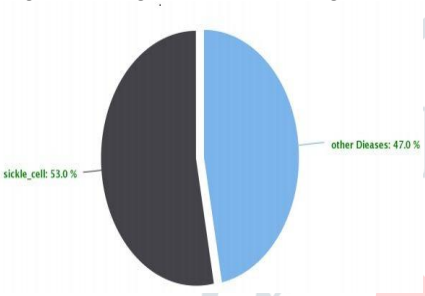

Fig 4. Percentage Of Sickle cell

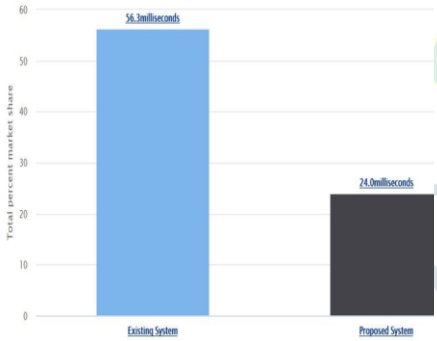

Fig. 5. System Comparison Graph

5. CONCLUSION

In This Application using data mining methods to discover the co-relation between the regular physical symptoms and the possible health risk given by the user or patients. Various and Efficient Machine learning algorithms are using to predict physical status user according to symptoms given and question answering. In our System patient get the Current Location and search the hospital, then results are given according to the current location of user/patients. User/patients gives symptoms and the system will predict the diseases and will recommend the medicines and will give the Online Doctor Appointment system. Our System design a feedback mechanism for doctors to fix classification and prediction result .Doctor collect prediction result through an user's side which will gather doctors' advice as new training data. Thus, our system could increase the execution and performance of prediction model automatically.

REFERENCES

[1] Guopeng Zhou ,Zhaoqian Lan, Wei Yan ,Yichun Duan "AI-assisted Prediction on Potential Health Risks with Regular Physical Examination Records", IEEE Transactions On Knowledge And Data Science ,18

[2] Rani B K, Govrdhan A, Srinivas K. "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks". International Journal on Computer Science & Engineering, 10.

[3] Wright A, Osheroff J,Sittig D, et al. "Grand challenges in clinical decision support". Journal of Biomedical Informatics, 08.

[4] Chang D C, Anderson J E. "Using Electronic Health Records for Surgical Quality Improvement in the Era of Big Data" [J]. Jama Surgery, 15.

[5] Petre R, Gheorghe M. "Integrating Data Mining Techniques into Telemedicine Systems" Informatica Economica Journal, 14.

[6] C. Hurtado, M. Mendoza and R. Baeza-Yates "Query recommendation using query logs in search engines," in Proc. Int. Conf. Current Trends Database Technol., 04.

[7] Tan G and Koh H C. Data mining applications in healthcare.[J]. Journal of Healthcare Information Management Jhim, 05.

[8] Kim J,Menon A K, Jiang X, et al. Detecting Inappropriate Access to Electronic Health Records Using Collaborative Filtering[J]. Machine Learning, 14.

[9] Accreditation Reports to Identify Risk Factors in Medication Management in Australian Residential Aged Care Homes [J]. Studies in Health Technology & Informatics, 17.

[10] Arnrich B, Lichte O, Netkeeper T W, et al. Evaluation of radiological features for breast tumor classification in clinical screening with machine learning methods [J]. Artificial Intelligence in Medicine, 05.

[11] Venkatesh S S, Conant E A, Song J H, et al. Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses.[J]. Academic Radiology, 05.

[12] West D, West V. Model selection for a medical diagnostic decision support system: a breast cancer

detection case.[J]. Artificial Intelligence in Medicine, 2000.
[13] Johnson M L, Gordon H S, Geraci J M, et al. Mortality after cardiac bypass surgery: prediction from administrative versus clinical data.[J]. Medical Care, 05.
[14] M.Durairaj, K.Meena, ―A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks, International Journal Of Innovative Technology & Creative Engineering ,2011.
[15] Gerald Tan and HianChyeKoh, ―Data Mining Applications in Healthcarr, journal of Healthcare Information Managemen.
[16] Kuo-Wei Hsu, JaideepSrivastava, PrasannaDesikan, ―Data mining for Healthcare Management, 11.
[17] ShwetaKharya, ―Using Data Mining Techniques ForDiagnosis and Prognosis of Cancer Disease, 12.
[18] V. Svátek and P. Nálevka "Improving Efficiency of Telemedical Prevention Programs through Data-mining on Diagnostic Data", 12.