

## Risk Prediction to Examine Health Status with Real and Synthetic Datasets

G. THIPPA REDDY<sup>1\*</sup>, APARNA SRIVATSAVA<sup>2</sup>, KURUVA LAKSHMANNA<sup>1</sup>,  
RAJESH KALURI<sup>1</sup>, SUDHEER KARNAM<sup>1</sup> and G. NAGARAJA<sup>1</sup>

<sup>1</sup>Assistant Professor, SITE, VIT University, India.

<sup>2</sup>Master of Computer Applications, SITE, VIT University, India.

\*Corresponding author E-mail: thippareddy.g@vit.ac.in

<http://dx.doi.org/10.13005/bpj/1309>

(Received: October 31, 2017; accepted: December 18, 2017)

### ABSTRACT

Now a days, every part of country try to take care of the health status of its public. There comes a process called health examination, which will predict health condition of the people. In this process the overall health records is merged into a single document and according to the data the prediction of risk will be calculated. Here we are using two types of data called real and synthetic, the real data comes under the data which we directly get through the hospital records and synthetic means the data which we have collect by ourselves. For the synthetic data we have to examine personally patient's health records. We may call the synthetic data as unlabeled because we don't have the exact records. The most important trial here is to predict the unlabeled one. This type of data-set is unique as it describes the person's health that is fluctuating i.e. good health to worst. In this paper we try to show the prediction of risk for the patient, whether the patient is good in health or they require some precaution. For this application we used an algorithm which is designed to detect the situation in process. Semi supervised is the main method for the entire application.


**Keywords:** Synthetic datasets, Health prediction, Examine records.

### INTRODUCTION

Data mining is a significant walk of learning divulgence plan which picks, enhance and illustrating enormous measure of data. It has transformed into a no matter how you look at it system in therapeutic science investigate. In restorative space it has expanded remarkable potential in finding the disguised cases from unlimited instructive accumulations. These cases are utilized for restorative conclusion to give better realizing which

can be valuable for the treatment. Portraying the unrefined helpful data is a slight dull undertaking, in light of the way that the data may make them miss or, on the other hand unessential data. Remedial decision candidly steady system assists therapeutic administrations specialists with settling on hospital opinion. To present Heterogeneous details about the health of the user and alert the user before they got in to ill status and it also give the detail about the health level of user and give detail about in which part they want to take more care. In this paper, we are



This is an  Open Access article licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits unrestricted Non Commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

trying to build an application that will early detect the health status of the patient. The process is one of the data mining method that we already discussed in our proposed work. The data which we got sometimes not sufficient to predict the entire situation, so that's why we are using each and every type of data. The health records from the hospital as well as the records which the user want to send us their information. For the information sharing purpose we provide our application. The existing system does not talk about the synthetic data. Mainly that system is dealing with the real data. So the application that we are building is unique. The previous system is based on classification techniques. Also it will provide some precaution to be taken. The algorithm that it used is SSL. The algorithm work as the process, it will classify the related data. SSL that gains from both real and synthetic information, and Positive and synthetic learning. A unique instance of SSL that is achieved by the real and synthetic information alone. Handles substantial and truly synthetic wellbeing information. These two techniques are implementing for double arrangement with pre-defined anti bin.

### Background

Utilizing this necessity, our application gives high administration proficiently. Programming necessities manage characterizing programming asset prerequisites and pre-essentials which should be introduced within the server which will give ideal working to the application. These necessities are large excluded in the product establishment bundle and should be introduced independently before the product is introduced. The most widely recognized arrangement of necessities characterized by any working framework or programming application are the substantial PC assets otherwise called equipment, equipment prerequisites rundown is frequently joined by an equipment similarity list (HCL), particularly if there should be an occurrence of working frameworks. The HCL records evaluate, perfect and some of the time incongruent equipment gadgets for a specific working framework. The accompanying sub-segments talk about the different parts of equipment necessities.

### Drawbacks

- Applications in healthcare only address binary classification problem.

- Multi-class classification problem with substantial unlabeled cases.

Through this paper we get the knowledge that we can predict future problems of the patient. There are various ways to predict the problem, with the help of graphs pattern we can predict. The first thing to do is to diagnosis patient at early stage. This technique will reduce the risk. For this type of system there must be provide some early prediction that the patient need some precaution. This paper gives us the brief idea about how to examine the risk at right time. The physicians use the interpret method that will easily shows the models. Here a data mining technique is used to diagnosis the result with this we can get the model pattern. This pattern will gives the way to treat patient at early stage.

The current wide selection of electronic therapeutic records presents awesome open doors and difficulties for information mining. The EMR information are generally transient, frequently boisterous, unpredictable and high dimensional. The paper builds a novel relapse structure for anticipating restorative hazard coming from EMR. Initially a theoretical perspective of EMR as a transient picture is developed to remove a various arrangement of elements. Other is the ordinal display that is connected for anticipating total or dynamic hazard. The difficulties are building a straightforward prescient model that will work with an expansive number of pitifully prescient elements but in meantime it should be steady against re-pattern varieties.

Cancer characterization is considered as the basic reason for patient-custom fitted treatment. Regular histological examination has a tendency to be problematic on the grounds that distinctive tumors may have comparative appearance. The latest innovation called microarray make respective treatment conceivable. Different machine learning strategies can be utilized to order cancer tissue tests in view of microarray information. In any case, couple of strategies can be richly embraced to create precise and dependable and additionally naturally interpretable tenets.

The capacity to foresee sharpness (patients' care needs), would give a capable instrument to

human services chiefs to dispense assets. Such estimations and expectations for the care procedure can be created from the immeasurable measures of human services information utilizing data innovation and computational insight systems. Strategic basic leadership and asset portion may likewise be upheld with various scientific enhancement models.

Determining of arrangement pictures in the prescription frequently depends on information named by a human master. As the labeling of Hospital information might be tedious so discovering methods for easing the naming expenses is basic for our capacity to consequently gain knowledge from this type of pictures. Through this paper we able to understood that machine learning approach can learn to enhanced double grouping models and the more proficiently by going through refining the twofold class data in the preparation stage with delicate marks that will shows unequivocally the human master expresses about the first class names.

**METHODOLOGY**

SHG on Health as a confirmation based hazard forecast way to deal with mining magnitude well being examination data set records. To deal with heterogeneity it investigates a Heterogeneous chart



**Fig. 1: ER diagram**

and huge amount of unlabeled information. This technique includes some special kind of strategy.

**Input and output process**

**Admin**

**Authentication**

Input: Give username and secret key to get consent for get to.

Output: Ended up plainly confirmed individual to demand and process the demand.

**Disease Record Maintain**

Input: Enter all the details regarding the Disease which is going to be stored in the database.

Output: Data saved with the input provided successfully in the database.

**User Record Analyze**

Input: Evaluate the record and compare the symptoms with the disease database.

Output: Clear ask to the user whether they are having symptoms regarding disease.

**User**

**Authentication**

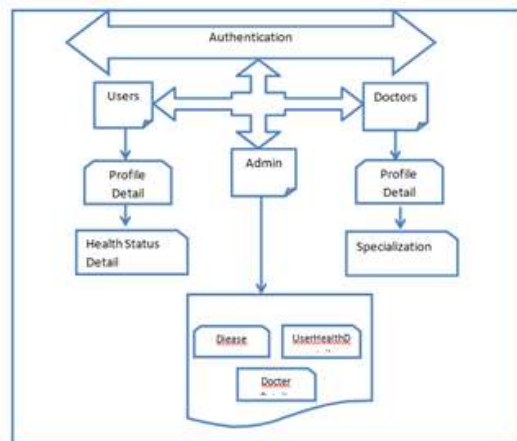
Input: Give the username along with secret key for access.

Output: Became authorized person, user get access to the process.

**User Profile**

Input: User can add or update its Profile information.

Output: Profile Updated Successfully.



**Fig. 2: System Architecture**

**Health Status Update**

Input: Enter their health details periodically.

Output: Health Status will get Updated Successfully.

**Technique used**

Graph drawn with heterogeneous details of various part of body and it analyze the detail with record set.

Step 1: Admin Create a record set Disease and their details

Step 2: User enter their health record in database

Step 3: Analyse the user record with Disease Record set

Step 4: Provide Consulting Doctor detail to the user. The overall process is explained through the ER model as shown in Fig 1.

**Proposed system architecture**

In our proposed system we are using prediction technique of the data mining. In this

technique the risk of the patient to get ill will be predicted. The system is based on the SHG design, this method contains semi-supervised process while in the existing system they are using supervised process. In this we are using real as well as synthetic data, by combining both data the system will give its prediction. There are various advantages as we have various data that includes every records, so the problem will be reduced.

The fig 2 is the system architecture, according to which we have user, admin and doctor as main module. The user has its sub module i.e. user profile and health status detail. Similarly in the doctor, its sub modules are doctor profile and specification. In admin module there is a disease, doctor and user health sub modules. All the above modules are combined together and related work will be done. The admin can access the user and doctor profile but user and doctor are allowed to view only their own profile. Admin can update its profile also it can update user's medical record. Like this the whole system will work with a graph based technique called semi supervised learning. Through this method the overall risk is predicted and according to which doctor will prescribe medicines. So with this method the risk of the patient health will reduce.

**Implementation**

This part describes the implementation of the paper. We will explain every point in wise with

Column Name	Data Type	Allow Nulls
ReqId	int	<input type="checkbox"/>
Name	nvarchar(50)	<input checked="" type="checkbox"/>
FatherName	nvarchar(50)	<input checked="" type="checkbox"/>
DateOfBirth	nvarchar(50)	<input checked="" type="checkbox"/>
Email	nvarchar(50)	<input checked="" type="checkbox"/>
Password	nvarchar(50)	<input checked="" type="checkbox"/>
Address	nvarchar(100)	<input checked="" type="checkbox"/>
Gender	nvarchar(50)	<input checked="" type="checkbox"/>
Country	nvarchar(50)	<input checked="" type="checkbox"/>
UserType	nvarchar(50)	<input checked="" type="checkbox"/>
Image	image	<input checked="" type="checkbox"/>

**Fig. 3: Database table**

Column Name	Data Type	Allow Nulls
BloodId	int	<input type="checkbox"/>
BloodGroup	nchar(10)	<input checked="" type="checkbox"/>
Donate	nvarchar(50)	<input checked="" type="checkbox"/>
Receive	nvarchar(50)	<input checked="" type="checkbox"/>

**Fig. 4: Blood table**

Column Name	Data Type	Allow Nulls
UniversalId	int	<input type="checkbox"/>
DiseaseName	nvarchar(50)	<input type="checkbox"/>
Symptoms	nvarchar(50)	<input type="checkbox"/>
Factors	nvarchar(50)	<input type="checkbox"/>
Infectors	nvarchar(50)	<input type="checkbox"/>
AffectedOrgans	nvarchar(50)	<input type="checkbox"/>
BasicStage	nvarchar(50)	<input type="checkbox"/>
AdvancedStage	nvarchar(50)	<input type="checkbox"/>
Precations	nvarchar(50)	<input type="checkbox"/>
Treatments	nvarchar(50)	<input type="checkbox"/>
Doctors	nvarchar(50)	<input type="checkbox"/>
Images	nvarchar(MAX)	<input type="checkbox"/>
Urls	nvarchar(MAX)	<input type="checkbox"/>

**Fig. 5: Centralized Reference Table**

Column Name	Data Type	Allow Nulls
DiseaseId	int	<input type="checkbox"/>
Disease	nvarchar(100)	<input checked="" type="checkbox"/>

**Fig. 6: Disease table**

Column Name	Data Type	Allow Nulls
SymptomsId	int	<input type="checkbox"/>
Symptom	nvarchar(100)	<input checked="" type="checkbox"/>

**Fig. 7: Symptoms table**



**Fig. 8: Graph on health status**

several snapshots. The database table is taken for our implementation.

**Database design Structure**

Database configuration is the solution toward delivering a point by point information picture of the database. This coherent information demonstrate contains all related intelligent and physical plan decisions and physical stockpiling guidelines expected to produce an outline in the Information explanation word that is used to make a database structure. A completely ascribed information contains point by point traits for every element.

The blood table is having blood Id, blood group, donate and receive field.

In the above fig 5, we have a reference table, this table will guide us directly to the doctor or symptoms as required

In fig 6 a disease table is there which have disease Id and type of disease records.

In fig 7, we have symptoms table it also has two parts symptoms Id and symptoms record. This table is required if doctor wants to see the patient's symptoms. When symptoms table is provided to the doctors then they will forward medicine to the relevant patients.

After the database table we will get the graph based on the SHG health algorithm.

**Future Enhancement**

In future work, In the future enhancement, Access level of data is deiced by the data provider. According to the level of data access beyond that level. The content that need to place in a multiple cluster is done by the content distribution over multiple cluster without place the entire content other than it place the reference of the content. It conserves the space memory wastage in data storage.

**CONCLUSION**

Extraction of well being examination information is generally testing because of its heterogeneity, characteristic commotion, and especially the huge amount of unlabeled information. In the project, we proposed a successful and adequate diagram algorithm called SHG Health, it will help to address these difficulties. Our prospective chart based characterization approach based on extraction of health records is having a couple of influence.

## REFERENCES

1. F. R. Institute, "Personal data in the cloud: A global survey of consumer attitudes," <http://www.fujitsu.com/downloads/SOL/fai/reports/fujitsu/personal-data-in-the-cloud.pdf>, 2010.
2. D. Quick and K. R. Choo, "Google drive: Forensic analysis of data remnants," *J. Network and Computer Applications*, **40**: pp. 179–193 (2014).
3. H. Chung, J. Park, S. Lee, and C. Kang, "Digital forensic investigation of cloud storage services," *Digital Investigation*, **9**(2), pp. 81–95 (2012).
4. D. Boneh and M. K. Franklin, "An efficient public key traitor tracing scheme," in *CRYPTO*, pp. 338–353 (1999).
5. D. Boneh, A. Sahai, and B. Waters, "Fully collusion resistant traitor tracing with short ciphertexts and private keys," in *EUROCRYPT*, pp. 573–592 (2006).
6. Z. Liu, Z. Cao, and D. S. Wong, "Traceable CP-ABE: how to trace decryption devices found in the wild," *IEEE Trans. Information Forensics and Security*, **10**(1), pp. 55–68 (2015).
7. D. Boneh and M. K. Franklin, "Identity-based encryption from the weil pairing," in *CRYPTO*, pp. 213–229 (2001).
8. A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *EUROCRYPT*, pp. 457–473 (2005).
9. V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *ACM Conference on CCS*, 2006, pp. 89–98.
10. R. Ostrovsky, A. Sahai, and B. Waters, "Attribute-based encryption with non-monotonic access structures," in *ACM Conference on Computer and Communications Security*, pp. 195–203 (2007).
11. C. Y. Wu, Y. C. Chou, N. Huang, Y. J. Chou, H. Y. Hu, and C. P. Li, "Cognitive impairment assessed at annual geriatric health examinations predicts mortality among the elderly," *Preventive Medicine*, **67**: pp. 28–34 (2014).
12. "Health assessment for people aged 75 years and older," [http://www.health.gov.au/internet/main/publishing.nsf/Content/mbsprimarycare\\_mbsitem75andolder](http://www.health.gov.au/internet/main/publishing.nsf/Content/mbsprimarycare_mbsitem75andolder), accessed: 2015-05-03.
13. "Health checks for the over-65s," <http://www.nhs.uk/Livewell/Screening/Pages/Checkover65s.aspx>, accessed: 2015-05-03.
14. L. Krogsbøll, K. Jørgensen, C. Grønhøj Larsen, and P. Gøtzsche, "General health checks in adults for reducing morbidity and mortality from disease (Review)," *Cochrane Database of Systematic Reviews*, **10** (2012).
15. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Data Mining and Knowledge Discovery*, **4**(4): pp. 1070–1093 (2015).
16. J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *Journal of the American Medical Informatics Association : JAMIA*, **20**(4): pp. 613–618 (2013).
17. H. Huang, J. Li, and J. Liu, "Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis," *Expert Systems with Applications*, **39**(3): pp. 2314–2320 (2012).
18. T. P. Nguyen and T. B. Ho, "Detecting disease genes based on semisupervised learning and protein-protein interaction networks," *Artificial Intelligence in Medicine*, **54**(1): pp. 63–71 (2012).
19. V. Garla, C. Taylor, and C. Brandt, "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management," *Journal of Biomedical Informatics*, **46**(5): pp. 869–875 (2013).
20. M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extraction of interpretable multivariate patterns for early diagnostics," *IEEE International Conference on Data Mining*, pp. 201–210 (2013).
21. T. Tran, D. Phung, W. Luo, and S. Venkatesh, "Stabilized sparse ordinal regression for medical risk stratification," *Knowledge and Information Systems*, pp. 1–28 (2014).
22. M. S. Mohhtar, S. J. Redmond, N. C. Antoniadis, P. D. Rochford, J. J. Pretto, J. Basilakis, N. H. Lovell, and C. F. McDonald, "Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data,"



- Artificial Intelligence in Medicine*, **63**(1): pp. 51–59 (2015).
23. J. M. Wei, S. Q. Wang, and X. J. Yuan, "Ensemble rough hypercuboid approach for classifying cancers," *IEEE Transactions on Knowledge and Data Engineering*, **22**(3): pp. 381–391 (2010).
23. E. Kontio, A. Airola, T. Pahikkala, H. Lundgren-Laine, K. Junttila, H. Korvenranta, T. Salakoski, and S. Salanterä, "Predicting patient acuity from electronic patient records." *Journal of Biomedical Informatics*, **51**: pp. 8–13 (2014).
24. Q. Nguyen, H. Valizadegan, and M. Hauskrecht, "Learning classification models with soft-label information." *Journal of the American Medical Informatics Association : JAMIA*, **21**(3): pp. 501–8 (2014).
25. G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus," *IEEE Transactions Knowledge and Data Engineering*, **27**(1): pp. 130–141 (2015).
26. L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q. Z. Sheng, and M. Sharaf, "Mining Personal Health Index from Annual Geriatric Medical Examinations," in 2014 IEEE International Conference on Data Mining, pp. 761–766 (2014).
27. S. Pan, J. Wu, and X. Zhu, "CogBoost: Boosting for Fast Costsensitive Graph Classification," *IEEE Transactions on Knowledge and Data Engineering*, **6**(1): pp. 1–1 (2015).
28. M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci, "A survey and analysis of Electronic Healthcare Record2005standards,"