Healthcare Data Mining: A Survey

Dr.R.Ramachandiran
Department of Information Technology
Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry
Email- ramachandiran@smvec.ac.in

Mr.K.S.Seetharaman
Department of Information Technology
Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry
Email- Seetharaman@smvec.ac.in

I.Deepika
Department of Information Technology
Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry
Email- deepikaa19990414@gmail.com

R.Neevethika
Department of Information Technology
Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry
Email- rngrk1998@gmail.com

Abstract- Due to the enormous amount of data available, there has been a great focus on the healthcare field. Not only efficient use of healthcare data make hospitals run more economically, but it has helped a lot in patient care and for saving lives of many. The services provided by healthcare is not the responsibility of medical sector but also of information technology. By having a good understanding of both historical and real time insights, healthcare analytics has the ability to help make improvements in all important areas before issues arise and help spot fatal medical issues in patient before they occur. In fact, data mining plays an active role in providing a consistent accuracy in predicting the diseases and its risk factors. Some of the data mining applications and techniques used in real world are discussed.

Keywords – Healthcare data mining, Machine Learning, Deep learning, Risk Prediction .

I. INTRODUCTION

One of the sector that greatly benefit from the adoption of data analytics is the healthcare sector. Hospitals make use of analytics software to make every information of their day to day running more efficient, more predictive and more intuitive. Healthcare data mining is the analysis and processing of data in the healthcare field in order to gain insights and support decision-making. From the key areas like medical expense, hospital data, patient behavior, and drugs, healthcare data mining can be used on both micro and macro range to effectively process operations, improve patient health, and lower overall expense. With the help of digitized electronic health records, patient histories and patterns can be determined more efficiently. Prediction of patient at risk can help isolate them and provide additional care and follow up treatment under crisis situation from chronic health problems. Physicians are given opportunity to provide corrective plans that decrease emergency visits. Monitoring those patients and providing respective care solutions is not possible without appropriate data, so the use of a Business Intelligence in healthcare is of great importance to protecting high-risk patients.

II. DATA MINING

Data mining is the process of analyzing the existing data to find patterns. According to Witten and Eibe[1], these trends should be "meaningful in that they offer some benefit, typically an economic advantage. "Data mining in the field of Healthcare is a gradually developing methodology that is used to identify a legitimate, useful relationships and patterns in the data that are too tedious and hard for human to understand. The existing unknown relationship and pattern in a database can be easily identified with a Data mining algorithms like classification, association, clustering, etc and with this we can build a Prediction Model. To identify the unknown patterns, we use data mining life cycle model. It includes the entire process from selecting data to exploiting data.

2.1 Applications Of Data Mining In The Field Of Healthcare –

Data mining supports all those interested in health care, including health insurers, healthcare companies, doctors and patients. Data mining can help identify insurance fraud and abuse for health care insurers. For Healthcare organizations, it supports Hospital Management system, Patient Tracking System and thereby forming a strong Customer relationship management. It can help predict the LOS, mortality rate, risk factors and rate of re-admission for patients and physicians. A doctor can choose the effective treatment available to the patient, with the above. According to a survey, 87% of the deaths in the US hospital have been prevented with the Predictive tool.

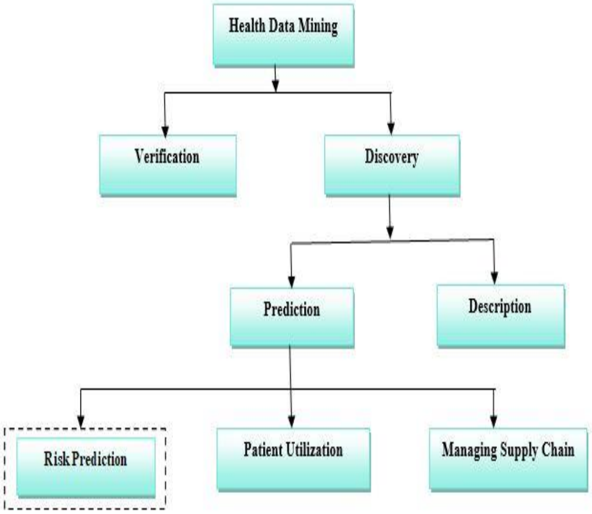

Figure1. Paradigms of Data mining

III. PREDICTIVE ANALYTICS

One of the major challenges in the field of healthcare is the quality of service and affordable cost for the service rendered. Quality treatment does not set the cost of care backfire. Predictive Analytics is in reality a major US data initiative. With a predictive Analytics tool, Physicians can only be a consultant and hospital readmission will be a problem for patients. So, the US health organizations try hard not to re-admit a patient. The care turns to Outcome from the pay-per-service program – driven payment system. Proper diagnosis and appropriate care should be done. In the end, it is important to record every single detail for further study. This function can be served by computer based knowledge. HPN, Heritage Prize Network was conducted by the US as an early prediction competition to boost healthcare, and the best algorithm was rewarded with a prize of $3,000,000. Predictive analytics allow for more efficient tracking and more cost-effectiveness for a safer environment.

3.1 The Significance Of Predictive Analytics In Healtcare–

To better understand the different possibilities of predictive analytics in health care, understanding the different ways in which health care can benefit from this discipline is first of all necessary. These include clinical management, such as overall business processes improvement; personal medicine to assist and improve diagnosis and treatment accuracy; and retrospective therapy and epidemiology to determine potential risk factors for public health

IV. RISK PREDICTION

Risk prediction is applicable to many issues in clinical medicine, public health, and epidemiology, and the expected risks of a specific disease or health result can be used to help patient, physician, health policymaker, and academic decision taking.

Table -1 illustrating potential applications of risk prediction models.

Domain	Question regarding risk	Decision based upon risk
Clinical Medicine	"Am I prone to heart attack"	Aspirin or no aspirin
Public health	"How many residents in public housing will have a heart attack this year?"	Defibrillator quantity and location in housing project
Epidemiology	"How many cardiac attacks do I expect from my clinical trial's control arm?"	Enroll more or fewer delegates

IV. SURVEY OF LITERATURE

The healthcare sector has produced huge volume of patient data in the recent days. Machine learning provides a way to search for trends and make data predictions. There are many recent works on data mining, with various types of digital data from patients. Deep learning has been used as an analytical way to tackle the complex problem of selecting apps. As stated in [7], in medical data, profound learning was applied to automatically pick and generate complex features from the input data. For example, Cheng et al. [8] suggested the use of Electronic Health Records (EHRs) as a deep learning-based prediction model. In this model, patient EHRs were represented as a multidimensional matrix including time and case.

Researchers used a four-layer convolution neural network (CNN) to derive characteristics and determine risks. In another study, Liang et al.[9] applied a deep learning model to the EHRs database to improve representations of features in making healthcare decisions. They suggested a method that would incorporate unsupervised as well as supervised learning. In particular, deep belief network (DBN) is applied to extract complex features, and then performed vector support machine (SVM). Nie et al.[10] suggested a deep learning scheme in their work to infer the potential diseases of people, considering the questions that people ask online. Further deep learning applications can be found in medical informatics and the fields of public health [11]–[15]. Different methods were explored to address the problem of data sparsity in medical data. For example, Wang et al.[16] proposed a high-order extension of the sparse logistic regression model, MulSLR (for Multilinear Sparse Logistic Regression), to predict clinical risk, process sparse and non-vector input data. In comparison to standard regressions of logistics, their method solved vectors of classification K. To assess patient acuity using incomplete, sparse and heterogeneous clinical data, Ghassemi et al. [17] proposed an approach that used the hyper-parameters of multi-task GP (MTGP) models to translate these clinical data into a new latent room. In this way, patients can be compared in the new hyper-parameter space, based on their similarity. Data could be viewed as time series data in this hyper-parameter space, and abstracted features can reflect dynamics of the series. This method was accepted to improve classification efficiency on ICU patient mortality prediction but the computational cost of this approach was very high. Lipton et al.[18] suggested a model approach using recurrent neural networks (RNNs) for the missing clinical data. Unlike traditional approaches that handle missing data in their approach by heuristic imputation, the authors modeled missingness as a feature. For missingness the proposed RNN can only use simple binary indicators. GRU-D [19], a deep learning model, was designed to exploit missing data patterns for successful imputation and performance-enhancing prediction. GRU-D has been developed using \the Gated Recurrent Unit (GRU), a recurrent neural network. GRU-D took masking and time interval as two representations of missing patterns and then incorporated them into an architecture of deep modeling. This model will capture the time series of the long-term temporal dependencies. Alternatively, the missing interest trends can be estimated to achieve better predictive performance. Despite the numerous efforts and successes of existing research in analyzing digital medical data, the analysis of medical data remains an important and demanding activity. Precise and effective risk prediction has always been an important topic which attracts the interests of many researchers

Table -1 Literature Survey Analysis

YEAR	TITLE	ANALYSIS
2018	Risk prediction of type II diabetes	Random forest algorithm uses several decision trees to train the samples, and combines weight of each tree to get the final results [21]
2018	A machine learning model to predict the risk of 30-day readmissions in patients with heart failure	The risk prediction model was created using deep unified networks (DUNs), a new mesh-like deep learning network framework designed to avoid overfitting [22].
2015	A comparison of models for predicting early hospital readmissions	The model is able to capture much richer structure by stacking multiple layers and learn substantially more complicated functions than a single layer neural network (DNN) [24]
2015	Predicting readmission risk with institution-specific prediction models	They experimented with methods of classification such as supporting vector machines and methods of prognosis such as the Cox regression.[23]
2014	Deep Learning for Healthcare Decision Making with EMRs	They have built up the framework for decision making using the multi-layer neural network [9]
2014	Clinical risk prediction with multilinear sparse logistic regression	They have proposed MulSLR: Multilinear Sparse Logistic Regression. MulSLR can be viewed as a high order extension of sparse logistic regression [16]
1999	Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data	A basic approach was developed using multiple regression models, which divided patients into small, moderate and high readmission risks [25]

IV. CONCLUSION

Data Mining is the process of finding interesting patterns out of huge amounts of data. Data Mining has a significant and increasing need for data analysis in the healthcare sector. Healthcare Data Mining has broadened its application to estimate patient risk levels in the area of healthcare. Many techniques in data mining include relationships, classifications, sequential patterns and clustering. The expertise gained through the application of techniques like data mining or machine learning can be useful in making successful healthcare decisions.

Various machine learning approaches for data analysis in healthcare are studied in this survey. It broadly analyses, discusses and classifies a range of techniques used in machine learning. Also analyzed the efficiency of the different machine learning algorithms

REFERENCES

[1] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, San Francisco, CA, USA, 2nd edition, 2005.
[2] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3):37, 1996..
[3] S. Purushotham, Y. Liu, and C.-C. J. Kuo, ”Collaborative topic regression with social matrix factorization for recommendation systems”, Proceedings of the 29th International Coference on International Conference on Machine Learning, 2012.
[4] Centers for Disease Control and Prevention. ”National Diabetes Statistical Report. Estimates of Diabetes and Its Burden in the United States”, 2014. [Online]. Available: https://www.cdc.gov/diabetes/pdfs/data/2014-reportestimates-of-diabetes-and-its-burden-in-the-united-states.pdf. [Accessed: 24- Jul- 2018].
[5] M. Petersen, ”Economic costs of diabetes in the U.S. in 2012”, Diabetes Care, vol. 36, no. 6, pp. 1797-1797, 2013. [6] K. E. Bergethon, C. Ju, A. D. Devore, N. C. Hardy, G. C. Fonarow, C. W. Yancy, P. A. Heidenreich, D. L. Bhatt, E. D. Peterson, and A. F. Hernandez, ”Trends in 30-Day Readmission Rates for Patients Hospitalized with Heart Failure: Findings from the Get with the Guidelines-Heart Failure Registry”, Circulation: Heart Failure, vol. 9, no. 6, 2016.
[7] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo and G. Yang, ”Deep Learning for Health Informatics”, IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 4-21, 2017.
[8] Y. Cheng, F. Wang, P. Zhang, and J. Hu, ”Risk Prediction with Electronic Health Records: A Deep Learning Approach”. Proceedings of the 2016 SIAM International Conference on Data Mining, 2016.
[9] Z. Liang, G. Zhang, J. Huang and Q. Hu, ”Deep learning for healthcare decision making with EMRs”, 2014 IEEE
International Conference on Bioinformatics and Biomedicine (BIBM), 2014.
[10] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang and T. Chua, ”Disease Inference from Health-Related Questions via Sparse DeepLearning”, IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, pp. 2107-2119, 2015.
[11] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, ”Distilling Knowledge from Deep Networks with Applications to Healthcare Domain”, Proceedings of the American Medical Informatics Assocation Annual Symposium (AMIA), 2016.
[12] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, ”Learning to diagnose with lstm recurrent neural networks”, arXiv preprint arXiv:1511.03677, 2015.
[13] S. Mehrabi, S. Sohn, D. Li, J. Pankratz, T. Therneau, J. Sauver, H. Liu and M. Palakal, ”Temporal Pattern and Association Discovery of Diagnosis Codes Using Deep Learning”, 2015 International Conference on Healthcare Informatics, 2015.
[14]R.Miotto,L.Li,B.KiddandJ.Dudley,”DeepPatient:AnUnsupervised Representation to Predict the Future of Patients
from the Electronic Health Records”, Scientific Reports, vol. 6, no. 1, 2016.
[15] H. Shin, Le Lu, L. Kim, A. Seff, J. Yao and R. Summers, ”Interleaved text/image Deep Mining on a large-scale radiology database”, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[16] F. Wang, P. Zhang, B. Qian, X. Wang and I. Davidson, ”Clinical risk prediction with multilinear sparse logistic regression”. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’14, 2014.
[17]M.Ghassemi,M.A.F.Pimentel,T.Naumann,T.Brennan,D.A.Clifton, P. Szolovits, M. Feng, ”A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data”, Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence, 2015.
[18] Z. C. Lipton, D. C. Kale, and R. Wetzell, ”Modeling Missing Data in Clinical Time Series with RNNs”, arXiv preprint arXiv:1606.04130, 2016.
[19] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, ”Recurrent Neural Networks for Multivariate Timef Series with Missing Values”, Scientific Reports, vol. 8, no. 1, 2016. [20] UCI Machine Learning Repository.
[21] Weifeng Xu,” Risk prediction of type II diabetes based on random forest model” Conference: 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics February 2017
[22] Golas SB1, Shibahara T,"A machine learning model to predict the risk of 30-day readmissions in patients with heart failure",2018 Jun 22
[23] Yu S1, Farooq F1, van Esbroeck "Predicting readmission risk with institution-specific prediction models", 2015.
[24] Joseph Futoma"A comparison of models for predicting early hospital readmissions", 2014
[25]Philbin EF1, DiSalvo TG"Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data",1999
[26] E.Poonguzhali ”A Extensive Survey on the Classification algorithms for Lymph Disease prediction” IJRAR May 2019, Volume 6, Issue 2.
[27] R.Saravanan “Heart disease prediction using ANN”, IJPAM,vol 119 Issue no 14 2018.