TO PREDICT THE CHANCE OF HEART ATTACK USING LOGISTIC REGRESSION AND PERFORMANCE IMPROVEMENT USING DIFFERENT CLASSIFIERS

[1]Vaishali Shinde, [2]Dr. Yogesh Kumar Sharma, [3]Harish Barapatre
JJTU University, Rajasthan Dept of Computer Scinece[1], JJTU University, Rajasthan
Dept of Computer Scinece[2], JJTU University, Rajasthan Dept of Computer Scinece[3]
vaishalishinde22@gmail.com[1], dr.yogeshkumar@yahoo.in[2], harishkbarapatre@gmail.com[3]

ABSTRACT

Educational Data Mining is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyse educational data in order to study educational questions. This paper surveys the most relevant studies carried out in this field to date. Firstly, it introduces EDM and describes the different groups of users, types of educational environments and the data they provide. It then goes on to list the most typical/common tasks in the educational environment that have been resolved through data mining techniques and finally some of the most promising future lines of research are discussed. With the overwhelming successes gained in Big Data analysis in the Business Industry, it is little wonder that there is a strong belief in the academia that these successes can be replicated in the Education Sector. As new findings and outcomes of research crop up daily, it is my belief that amongst these successes potentially identifiable, prediction of students' academic performance can have strong positive influences in knowledge management and delivery in education thereby adding more quality to the learning experience.

Keywords— Data Mining, Student Database Mapping, Prediction analysis, Educational dynamics of behavior

INTRODUCTION

Health care is coming to a new era where the abundant biomedical data are playing more and more important roles. In this context, for example, precision medicine attempts to 'ensure that the right treatment is delivered to the right patient at the right time' by taking into account several aspects of patient's data, including variability in molecular traits, environment, and electronic health records (EHRs). The large availability of biomedical data brings tremendous opportunities and challenges to health care research. In particular, exploring the associations among all the different pieces of information in these data sets is a fundamental problem to develop reliable medical tools based on data-driven approaches and machine learning.

A risk prediction model aims to predict the probability or risk of a condition or event among individuals, or occasionally groups, based on a combination of known or measured characteristics. Risk prediction tools are the means by which risk prediction models, scores or algorithms are implemented in clinical practice. Numerous risk tools are now available, which predict either current or future risk of a various diagnosis. In these works we have the potential to improve patient outcomes through enhancing the consistency and quality of clinical decision-making, facilitating equitable and cost-effective distribution of finite resources and encouraging behaviour change.

Recently, Figure no-1 shows the data analyst system with the increasing availability of a large volume of electronic health record (EHR) data, there is a gradual attention to use data-driven approaches to construct efficient tools for clinical prediction. Electronic health record (EHR) data from millions of patients are now routinely collected across diverse healthcare institutions. They consist of heterogeneous data elements, including patient demographic information, diagnoses, laboratory test results, medication prescriptions, clinical notes, and medical images. However, it is challenging to create accurate analytic models from EHR data, because of data quality, data and label availability and heterogeneity of data types.

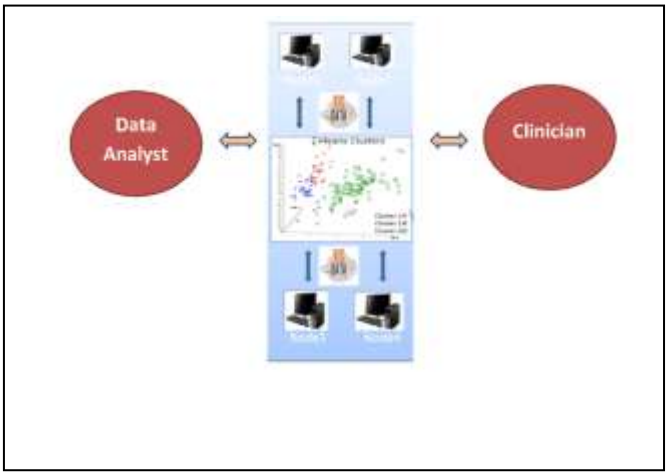

Fig-1: Data Analyst System

Traditional health analytics modelling often depends on labour intensive efforts, such as expert-defined phenotyping and ad-hoc feature engineering. The resulting models generally have limited generalizability across datasets or institutions. A common approach in biomedical research is to have a domain expert to specify the phenotypes to use in an ad hoc manner. However, supervised definition of the feature space scales poorly and misses the opportunities to discover novel patterns. Alternatively, representation learning methods allow to automatically discovering the representations needed for prediction from the raw data.

Deep learning methods are representation-learning algorithms with multiple levels of representation, obtained by composing simple but nonlinear modules that each transforms the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. Deep learning models demonstrated great performance and potential in computer vision, speech recognition and natural language processing tasks. Given its demonstrated performance in different domains and the rapid progresses of methodological improvements, deep learning paradigms introduce exciting new opportunities for biomedical informatics.

Until the last few years, most of the techniques for analysing rich EHR data were based on traditional machine learning and statistical techniques such as logistic regression, support vector machines (SVM), and random forests. Recently, deep learning techniques have achieved great success in many domains through deep hierarchical feature construction and capturing long-range dependencies in data in an effective manner. Given the rise in popularity of deep learning approaches and the increasingly vast amount of patient data, there has also been an increase in the number of publications applying deep learning to EHR data for clinical informatics tasks which yield better performance than traditional methods and require less time-consuming pre-processing and feature engineering.

The contribution of the proposed research is to design a machine-learning-based medical intelligent decision support system for the diagnosis of heart disease. In the present study, various machines learning predictive models such as logistic regression, k-NN, ANN, SVM, decision tree, Naive Bayes, and random forest have been used for classification of people with heart disease and healthy people. 0ree feature selection algorithms, Relief, minimal redundancy- maximal-relevance (Mr MR), Shrinkage and Selection Operator (LASSO), were also used to select the most important and highly correlated features that great influence on target predicted value. Cross-validation methods like k-fold were also used. In order to evaluate the performance of classifier, various performance evaluation metrics such as classification accuracy, classification error, specificity, sensitivity, Matthews' correlation coefficient (MCC), and receiver optimistic curves (ROC) were used. Additionally, model execution time has also been computed. Moreover, data pre-processing techniques were applied to the heart disease dataset. 0e proposed system has been trained and tested on Cleveland heart disease dataset, 2016. UCI data-mining repository the dataset of Cleveland heart disease is available online.

REVIEW OF LITERATURE

Review of literature refers to those research paper, publication, journals and articles which include finding of a researcher which can be theoretical or statically in nature on one particular topic, it is a secondary source of information and does not take into view any new finding. Here the researcher has taken the review of literature in India and abroad from various books and journals:

Wang, H [2014] he explained a comprehensive medical lexicon which supports the automatic lexical tagging process is indispensable. In this study, he manually extracted and encoded lexemes and their semantic classes in the clinical pathway specification published by the host hospital and merged them into the medical lexicon described in as our prototype lexicon.

LeCun Y [2015] they explained Deep learning methods are representation-learning algorithms with multiple levels of representation, obtained by composing simple but nonlinear modules that each transforms the representation at one level into a representation at a higher, slightly more abstract level. Deep learning models demonstrated great performance and potential in computer vision, speech recognition and natural language processing tasks

S. Bandyopadhyay [2015] he suggested the unique features and challenges of EHD, including missing risk factor information, non-linear relationships between risk factors and cardiovascular event outcomes, and differing effects from different patient subgroups, demand novel machine learning approaches to risk model development. Also, he explained a machine learning approach based on Bayesian networks trained on EHD to predict the probability of having a cardiovascular event within five years. In such data, event status may be unknown for some individuals as the event time is right-censored due to disenrollment and incomplete follow-up. He described how to modify both modelling and assessment techniques to account for censored observation times. They showed that the approach can lead to better predictive performance than the Cox proportional hazards model.

Liu et al. [2015] they used a four-layer CNN to predict congestive heart failure and chronic obstructive pulmonary disease and showed significant advantages over the baselines.

Choi E [2015] they incorporated medical interventions in the model to dynamically shape the predictions. DeepCare was evaluated for disease progression modelling, intervention recommendation and future risk prediction on diabetes and mental health patient cohorts. RNNs with gated recurrent unit (GRU) were used to develop Doctor AI, an end-to-end model that uses patient history to predict diagnoses and medications for subsequent encounters. The evaluation showed significantly higher recall than shallow baselines and good generalizability by adapting the resulting model from one institution to another without losing substantial accuracy.

Gulshan [2016] he has used CNNs to identify diabetic retinopathy in retinal funds photographs, obtaining high sensitivity and specificity over about 10 000 test images with respect to certified ophthalmologist annotations. CNNs also obtained performances on par with 21 board-certified dermatologists on classifying biopsy-proven clinical images of different types of skin cancer over a large data set of 130 000 images (1942 biopsy-labelled test images).

NEED OF THE STUDY

When a patient is admitted to a hospital, there are two commonly asked questions: ''what is happening?" and ''what happens next?" The first question is about illness diagnosis, the second is about predicting future medical risk. Whilst there are a wide array of diagnostic tools to answer the first question, fewer technologies address the second. Traditionally, the prognostic question may be answered by experienced clinicians who have seen many patients or by clinical prediction models with well-defined risk factors. But both methods are expensive and restricted in availability. Modern electronic medical records (EMRs) promise a fast and cheap alternative. An EMR contains the history of hospital encounters, diagnoses, interventions, lab tests and clinical narratives. The wide adoption of EMRs has led to recent research to build predictive models from this rich data source. Answering prognostic inquiries necessitates modelling patient level temporal healthcare processes. Effective modelling must address four open challenges:

1. Long-term dependencies in healthcare: the future illness and care may depend critically on historical illness and interventions. For example, the onset of diabetes in middle age remains a risk factor for a person's remaining life; cancers may recur after years; and a previous surgery may prevent certain future interventions.
2. Representation of admission information: an admission episode consists of a variable-size discrete set containing diagnoses and interventions.
3. **3.** Episodic recording and irregular timing: medical records vary greatly in length, are inherently episodic in nature and irregular in time. The data is episodic because it is only recorded when the patient visits the hospital and undergoes an episode of care. The episode is often tightly packed in a short period, typically ranging from a day to two weeks. The timing of arrivals is largely random.
4. **4.** Confounding interactions between disease progression and interventions. Existing methods are poor in handling such complexity. They inadequately model variable length and ignore the long-term dependencies. Temporal models based on the Markovian assumption are limited to model temporal irregularity and have no memory, and thus they may completely forget previous major illness given an irrelevant episode.

OBJECTIVE

The main objectives of this study are:

1. To develop robust prediction models that can effectively handle high dimensional heterogeneous EHR data.
2. To accurately classify different clinical risks levels based on the acquired EHR data.
3. To develop timely and appropriate intervention strategies to those at high risk levels.
4. To estimate the chance of a unfavourable major event (such as death) during patients' hospitalizations

PROPOSED SYSTEM

The proposed system architecture divided into five stages including:

Pre-processing of dataset

Feature Based selection

Cross validation with Reference method

Machine learning classifiers

Performance evaluation methods Using Different Classifier

Figure 3 shows the architecture of the proposed system which has been developed with the objective to classify people with heart disease level and healthy people. The performances of different machine learning predictive models for heart disease diagnosis on full and selected features were tested. Feature selection algorithms such as Relief, mRMR, and LASSO were used to select important features, and on these selected features, the performance of the different classifiers level with accuracy was calculated. The level and heart disease dataset have been implemented in several studies [13] and is used in our study. The popular machine learning classifiers logistic regression, K-NN, ANN, SVM, DT, and NB were used in the system. The model's validation and performance

evaluation metrics were computed. The stage of pre-processing of data level is necessary for efficient main representation of data category and machine learning tools which should be tested, classified and trained using evaluation classifier in effective manner. The main task of Pre-processing techniques is to remove of missing values, standard scalar, and Minmax Scalar which have been applied to the large dataset for effective use in the best classifiers. The standard scalar which ensures that every main feature has the mean ZERO and variance ONE, bringing all features to the same coefficient. Similarly, in Minmax Scalar shifts the data such that all features are between ZERO and ONE. The missing values feature row is just deleted from the dataset.

A. *Machine Learning Classifiers.*

In order to categories the heart attack patients and healthy people, machine learning classification algorithms are used to detect the heart attack level like medium low and high main popular classification algorithms and their theoretical background are discussed briefly in this section.

The aim of this study is to identify the most significant predicators of heart diseases and predicting the overall risks by using logistic regression. Thus, the following binary logistic library are used for the prediction of heart attack dieses

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import scipy.stats as st
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.metrics import confusion_matrix
import matplotlib.mlab as mlab
```

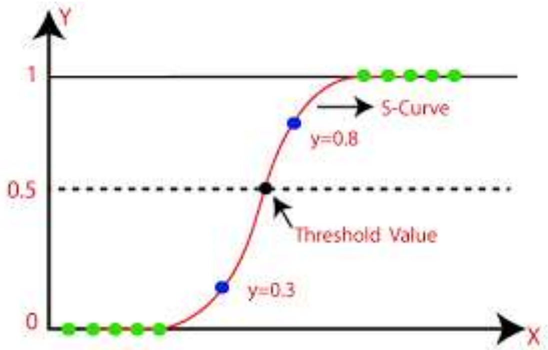

Fig-1: Max and Min Threshold Graph

The maximum and minimum threshold value graph is shown in figure-1. The Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

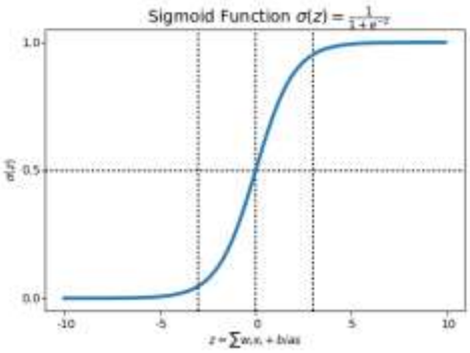

Fig-2 Sigmoid Graph Function

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

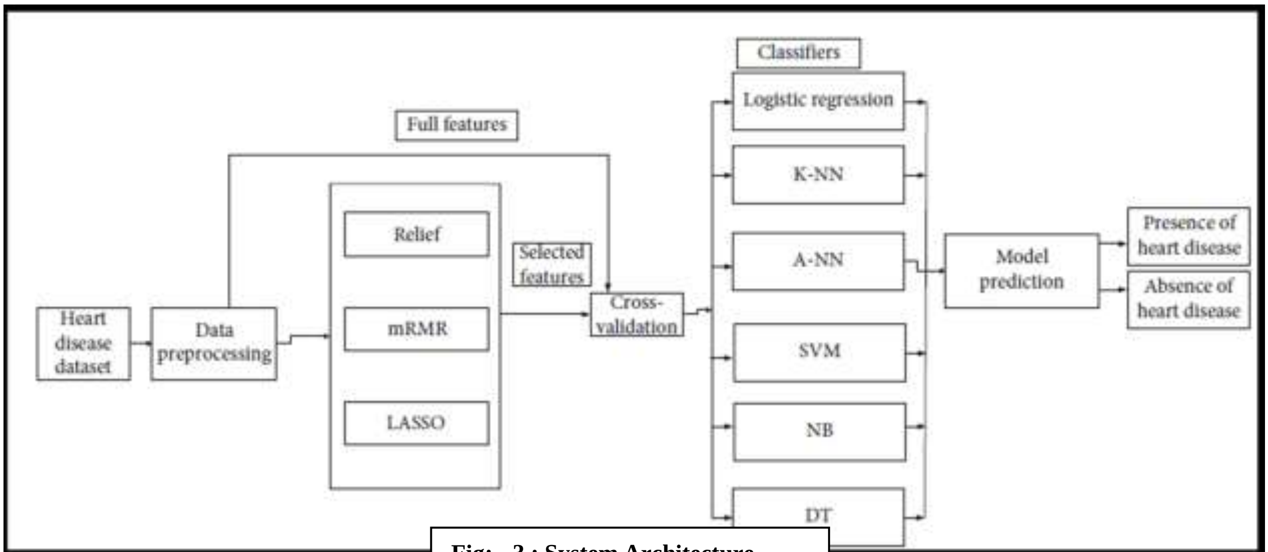

Fig: - 3 : System Architecture

S. no.	Feature name	Feature code	Description	Domain of values (min-max)
1	Age	AGE	Age in years	30 < age < 77
2	Sex	SEX	Male = 1 Female = 0	1 0
3	Type of chest pain	CPT	1 = atypical angina 2 = typical angina 3 = asymptomatic 4 = nonanginal pain	1 2 3 4
4	Resting blood pressure	RBP	mm Hg admitted at the hospital	94–200
5	Serum cholesterol	SCH	In mg/dl	120–564
6	Fasting blood sugar >120 mg/dl	FBS	Fasting blood sugar >120 mg/dl (1 = true; 0 = false)	1 0
7	Resting electrocardiographic results	RES	0 = normal 1 = having ST-T 2 = hypertrophy	0 1 2
8	Maximum heart rate achieved	MHR	—	71–202
9	Exercise-induced angina	EIA	1 = yes 0 = no	0 1
10	Old peak = ST depression induced by exercise relative to rest	OPK	—	0–6.2
11	Slope of the peak exercise ST segment	PES	1 = up sloping 2 = flat 3 = down sloping	1 2 3
12	Number of major vessels (0–3) colored by fluoroscopy	VCA	—	0 1 2 3
13	Thallium scan	THA	3 = normal 6 = fixed defect 7 = reversible defect	3 6 7

Fig-4 : Input Feature Training data set with min and max value

Logistic regression is used when you have a classification problem - yes/no, or a number between 1 and 10 representing an answer on a survey question. The logistic function looks like a big S and will transform any value into the range 0 to 1. This is useful because we can apply a rule to the output of the logistic function to snap values to 0 and 1 (e.g. IF less than 0.5 then output 1) and predict a class value. Because of the way that the model is learned, the predictions made by logistic

also be used as the probability of a given data instance belonging to class 0 or class 1. Logistic regression is one best prediction model as compared to another classifier. In Figure-4 shows the different value with their code also we have defined the maximum and minimum range of each value. The position of SVM is shown in figure-3 that SVM used a maximum margin strategy that transformed into solving a complex quadratic programming problem. Due to the high performance of SVM in classification, various applications widely applied it. The NB is a classification supervised learning algorithm. It is based on conditional probability theorem to determine the class of a new feature vector. The NB uses the training dataset to find out the conditional probability value of vectors for a given class. After computing the probability conditional value of each vector, the new vectors class is computed based on its conditionality probability. NB is used for text-concerned problem classification.

ALGORITHM

Stacked Denoising Auto-Encoder Algorithm

Stacked Denoising Auto-Encoder (SDAE), one of the most extensively investigated deep learning architectures, is employed in this study. SDAE is a symmetrical neural network, and mainly used for learning the features from dataset in an unsupervised fashion. Typically, each Denoising Auto Encoder in SDAE will be trained to reconstruct a clean "repaired" input from a corrupted version of it. The SDAE is useful to learn a hierarchy of features in a greedy layer-wise unsupervised model. The learning process starts to train the first auto-encoder by optimizing the loss function with the original input data to learn the first hidden representation layer. After that, the learned hidden layer is used as the input data for training the next auto-encoder to generate higher-level representations, and this process is repeated with K times, where K is the number of hidden layers. Since deep learning architectures, such as SDAE, deal with data abstraction and representation, it is quite likely to be suitable for analysing raw data presented in different formats and/or from different sources.

Supervised fine Tuning

Once the pretraining is complete, we will append a soft max regression layer on the top of the reconstructed feature representation layer to construct a deep neural network, named regularized stacked denoising auto-encoder with soft max regression model (RSDAE-SM), and the use this RSDAE-SM to perform clinical risk prediction task. Specifically, we fine tune the constructed RSDAE-SM using back propagation by minimizing the cross-entropy loss using for the soft max layer.

CONFUSION MATRICS

We have considered the following formula for improved the performance of the complete system. The matrix showed the complete relation between the TP, FP, TN, FN. And their relation. We used confusion matrix, every observation in the testing set is predicted in exactly one box. It is 2 × 2 matrix because there are 2 repose classes. Moreover, it gives two types of correct prediction of the classifier and two types of classifier of incorrect prediction.

Accuracy: accuracy shows the overall performance of the classification system which shown in figure-5 (1).

Model Error: it is the overall incorrect classification of the classification model which is calculated by using the formula (2) in figure no-5.

Sensitivity: it is the ratio of the recently classified heart patients to the total number of heart patients. The sensitivity of the classifier for detecting positive instances is known as "true positive rate." Is mentioned in figure 5 reference 3.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad \text{.................................... (1)}$$

$$\text{error} = \frac{FP + FN}{TP + TN + FP + FN} \times 100\% \quad \text{....................................(2)}$$

$$\text{Sensitivity (Sn)/recall/true positive rate} = \frac{TP}{TP + FN} \times 100\% \quad \text{.............. (3)}$$

$$\text{specificity (Sp)} = \frac{TN}{TN + FP} \times 100\% \quad \text{..................................(4)}$$

$$\text{precision} = \frac{TP}{TP + FP} \times 100\% \quad \text{................................(5)}$$

Fig-5: Performance Evaluation term

Specificity: a diagnostic test is negative and the person is Healthy which can be calculate by using the formula - 4. **Precision:** the equation of precision is given as in figure-5

RESULT DISCUSSION

The overall performance of system and prediction of heart attack range using different classifier with their performance is shown in figure-6

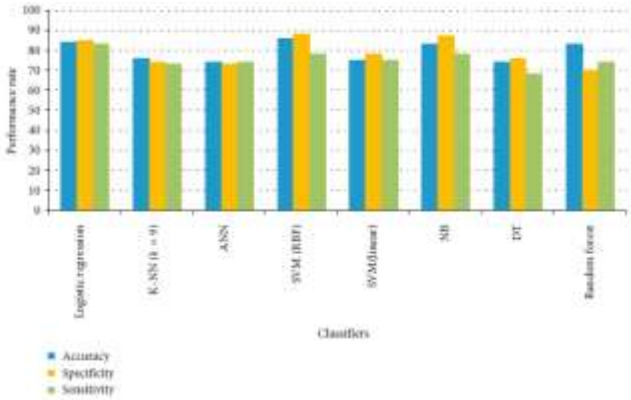

Fig-6: Performance Graph

Logistic regression algorithm gives 89% accuracy with 10 features and 6-fold as compared to other algorithm and prediction performance time is very high as compared to SVM algorithm. Our model predicts the chance of heart attack level that is low medium and high shows the accurate prediction.

CONCLUSION

In this research study, a machine-learning based predictive system was proposed for the diagnosis of heart disease which shows different level called low medium and high-risk prediction. The system was tested on risk heart disease dataset. Different classifiers such as logistic regression, K-NN, ANN, SVM, NB, DT, and random forest were used prediction model that is Logistic Regression used to select the important features. The K-fold cross-validation method was used in the system for validation. In order to check the performance of classifiers, different evaluation metrics were also adopted. The feature selection algorithms select important features that improve the performance of classifiers in terms of classification accuracy, specificity, and sensitivity and reduced the computation time of algorithms. The classifiers logistic regression with 10-fold cross-validation showed best accuracy 89% when selected by FS value. Due to the good performance of logistic regression with 10 cross fold, it is a better predictive system in terms of accuracy.

REFERENCE

1. **K. Goulianasl, M, . Adu-opoulost (1997)** "Structured Artificial Neural Notworks for Fast Batch LMS Algorithms", Neural ,Parallel& Scientific Computations, 54y562.
2. Yoav Freund and Robert E (1997), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", SchapireAT6T Labs, 180 Park Avenue, Florham Park, New Jersey 07932 Received December 19, 1996 , journal of computer and system sciences 55, 119139 (1997).
3. Friedman J, Hastie T, Tibshirani R. (2000) "Additive logistic regression: a statistical view of boosting", Ann Stat. 1998; 28:2000.
4. Welling, M., Rosen-Zvi, M., & Hinton, G. E. (2005) "Exponential family harmoniums with an application to information retrieval" In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), Advances in neural information processing systems, 17 (pp. 1481–1488). Cambridge, MA: MIT Press.

5. Steven Walczak (2005) "Artificial Neural Network Medical Decision Support Tool: Predicting Transfusion Requirements of ER Patients" IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 9, NO. 3, SEPTEMBER 2005.
6. Geoffrey E. Hinton (2006), "A Fast Learning Algorithm for Deep Belief Nets", Singapore 117543, Neural Computation 18, 1527–1554 (2006) C _ 2006 Massachusetts Institute of Technology, Department of Computer Science, University of Toronto, Toronto, Canada M5S 3G4.
7. Marc'Aurelio Ranzato, et al.(2007), "Unsupervised learning of invariant feature hierarchies with applications to object recognition", In: Proc. of Computer vision and Pattern Recognition Conference, 2007.
8. Meystre, S.M. (2008) "Extracting information from textual documents in the electronic health record: A review of recent research". Yearb. Med. Inform. 2008, 35, 128–144.
9. Bengio Y.(2009), "Learning deep architectures for AI", Found Trends Mach Learn 2009;2:1–127.
10. Wu, Stewart, W.F. (2010), "Prediction modelling using EHR data: Challenges, strategies, and a comparison of machine learning approaches", Med. Care, 2010, 48, S106–S113.
11. Bengio Y (2013) "Representation learning: a review and new perspectives". IEEE Trans Pattern Anal Mach Intell 2013;35:1798–828.
12. Wang, H., Feng, K. Liu, L. (2014), "Extracting important information from Chinese operation notes with natural language processing methods". J. Biomed. Inform. 2014, 48, 130–136.
13. Duan H. (2015) "Lexical Characteristics Analysis of Chinese Clinical Documents", In Proceedings of the 2015 7th International Conference on Information Technology in Medicine and Education (ITME), Huangshan, China, 13–15 November 2015; pp. 121–125.
14. Liu C, Wang F, Hu J, et al.(2015), "Risk prediction with electronic health records: a deep learning approach" In: ACM International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 2015, 705–14.
15. LeCun Y, Bengio Y, Hinton G. (2015), "Deep learning for EHR" Deep learning Nature 2015;521:436–44.
16. Choi E, Bahadori MT, et al. (2015), "Doctor AI: predicting clinical events via recurrent neural networks", arXiv 2015. http://arxiv.org/abs/1511.05942v11.
17. S. Bandyopadhyay, et al. (2015), "Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data", Data Mining and Knowledge Discovery, submitted (2014) 2015, 29(4):1033-1069.
18. Z. Huang, et al. (2015) "A probabilistic topic model for clinical risk stratification from electronic health records", Journal of Biomedical Informatics, 58:28-36.
19. Danqing Hu (2016) "Utilizing Chinese Admission Records for MACE Prediction of Acute Coronary Syndrome".
20. D. Hu, et al. (2016), "Utilizing Chinese Admission Records for MACE Prediction of Acute Coronary Syndrome", International Journal of Environmental Research and Public Health, 13(9):912, 2016.
21. Pham T, Tran T, Phung D, et al. (2016) "DeepCare: a deep dynamic memory model for predictive medicine" arXiv 2016. https://arxiv.org/abs/1602.00357.

22. J. E and Y. N. (2016) "Electronic Health Record Adoption and Use among Office-based Physicians" in the U.S., by State: 2015 National Electronic Health Records Survey." The Office of the National Coordinator for Health Information Technology, Tech. Rep., 2016.
23. Farhan W, Wang Z, Huang Y, et al. (2016), "A predictive model for medical events based on contextual embedding of temporal sequences". J Med Internet Res 2016;4:e39.
24. Gulshan V, Peng L, Coram M, et al. (2016), "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs", JAMA 2016; 316:2402–10.
25. Cheng J-Z, Ni D, Chou Y-H, et al.(2016), "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans", Sci Rep 2016;6:24454.
26. Miotto R, Li L, Kidd BA, et al.(2016), "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records", Sci Rep 2016;6:26094.
27. Riccardo Miotto and Fei (2017) "Deep learning for healthcare: review, opportunities and Challenges", Dudley , Briefings in Bioinformatics, 2017, 1–11.
28. Benjamin Shickel & Patric k J. Tighe (2017) "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis" 2168-2194 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee. org/publications_ standards/publications/rights/index.html for more information.
29. Z. Huang, et al.(2017), "MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records", Journal of Biomedical Informatics, 66:161-170, 2017.
30. P. Y. Wu, et al. (2017) "Omic and Electronic Health Record Big Data Analytics for Precision Medicine" IEEE Transactions on Biomedical Engineering, 64(2):263-273, 2017.