International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

A NOVEL LEARNING APPROACH FOR CLINICAL RISK PREDICTION DATA OF ACS USING EHR

Mr. Chandrabhan S. Jadhao[1], Mr. Mahesh Pimpalkar[2]

[1,2]Department of Computer Engineering SES's Yadavrao Tasgaonkar Institute Of Engineering And Technology, Karjat, Maharashtra, India

---

Abstract - *Acute coronary syndrome (ACS) is a term used to describe a range of conditions associated with impulsive, condensed blood flow to the heart.ACS is the most common type of coronary artery disease (CAD). Every year, CAD and ACS together account for approximately 7 million deaths, , and about 30% people are at risk of having ACS in their lifetime. Clinical risk prediction of ACS is significant for early intervention and treatment. Existing ACS risk scoring models are based mainly on a small set of hand-picked risk factors and often dichotomize predictive variables to simplify the score calculation. The last decade has seen an explosion in the amount of digital information stored in electronic health records (EHRs). Over the same period, the machine learning community has seen advances in the field of Deep Learning. In this aspect, we survey the current work proposal on applying Deep Learning to clinical tasks based on EHR data.This work proposes a regularized "stacked denoising auto-encoder" (SDAE) model to stratify clinical risks of ACS patients from a large volume of "electronic health records" (EHR). To capture characteristics of patients at similar risk levels, and preserve the discriminating information across different risk levels, two constraints are added on SDAE to make the reconstructed feature representations contain more risk information of patients, which contribute to a better clinical risk prediction result.*

Key Words: SDAE, EHR, coronary artery disease, acute coronary syndrome,deep learning.

1. INTRODUCTION

"Electronic health record" (EHR) data from millions of patients are now regularly collected across diverse healthcare institutions. They consist of heterogeneous data elements, including patient demographic information, diagnoses,test results, medication prescriptions, clinical notes, and medical images. However, it is challenging to create accurate analytic models from EHR data, because quality of data, data and label availability, and diversity of data types. Traditional health analytic often depends on labor intensive efforts, such as expert-defined typing and ad-hoc feature engineering. The resulting models often have limited generalizability across datasets or institutions. Deep learning has had a profound impact in various data analytic applications, such as image classification, speech recognition, computer vision, and natural language processing. It has changed the data analytic modelling paradigm from expert-driven feature engineering (EDFE)to data-driven feature construction (DDFC). Over the couple of years, an increasing body of literature confirmed the success of feature construction using deep learning methods (ie. models with multiple layers of neural networks). Interest in deep learning for healthcare has grown for two reasons. First, for healthcare researchers, deep learning models yield better performance in many tasks than traditional machine learning methods and require less manual feature engineering. Second, large and complex datasets (eg., longitudinal event sequences and continuous monitoring data) are available in healthcare and enable training of complex deep learning models. However EHR data also introduce many interesting modeling challenges for deep learning research. Acute coronary syndrome (ACS) is a term used to describe a range of conditions associated with sudden, reduced blood flow to the heart, including ST-elevation myocardial infarction (STEMI), non- ST-elevation myocardial infarction (NSTEMI), and unstable angina (UA). ACS is the most common type of coronary artery disease (CAD). Every year, CAD and ACS together account for approximately 7 million deaths, accounting for around half of the global burden, and about 30% people are at risk of having ACS during their lifetime. Thus Clinical risk prediction of ACS is important for early intervention.

With increasing availability of a large volume of electronic health record (EHR), there is a gradual attention to use data-driven approaches to form efficient tools for ACS prediction. While primarily designed for archiving patient information and performing administrative health care tasks like billing, many researchers have found secondary use of these records for various clinical informatics applications. Over the same period, the machine learning community has seen widespread advances in the field of deep learning. In this aspect, we survey the current project proposal on applying deep learning to clinical tasks based on EHR data.

2. EXISTING SYSTEM

2.1 Clinical Risk Prediction

Clinical risk prediction models are increasingly being used in health care for a widespread range of applications. In primary care, they may be used to aim interventions by classifying patients with higher risk of diseases such as ACS. In practice, a number of studies have been emerged with authenticated models to enable clinicians to reliably identify patients at low risk, medium risk, and high risk for ACS [12]. For example, the "GRACE risk scoring model", as one of the most popular and accepted risk scores of ACS, was developed to predict clinical risk score of individual patients [3]. It is notable that models along this line have been estimated using a small set of specially chosen patient features from highly-stratified cohorts. In consequences, they just account for a small number of hand-picked risk factors and are in fragments. Recently, many advanced data mining algorithms have been introduced for clinical risk prediction, with the widespread adoption of electronic health records (EHR) [6, 13, 19, 20]. The EHR typically records a diverse set of clinical information, including patient demographics, symptoms, laboratory test results, and treatment behaviors, etc. It, therefore, provides a comprehensive source for clinical risk prediction. Many data mining algorithms, such as decision trees [11], Bayesian networks [12], and fuzzy inference systems [15], etc., have been proposed to explore the potential of EHR data for clinical risk prediction. For example, Tay et al. presented a novel neural inspired algorithm for risk prediction by using EHR . Karaolis et al., applied the C4.5 decision tree algorithm to extract essential risk factors of coronary heart events from EHR data [14]. In [19], a hybrid model was developed for automatically identifying risk factors of heart disease in patient EHR.

2.2 Support Vector Machine

Another existing systems use Support vector machine (SVM). SVMs are less prone to over fitting and need less amount of memory. They have been proved to perform better for text classification in EHRs and pattern recognition in echocardiography imaging to stratify CV risk and helping physicians make decisions [8, 9]. In addition, an SVM is feasible for nonlinear data sets or large and complex datasets, such as "omic" data, because kernel functions, a shortcut to expedite learning process, can be applied to SVMs to outperform their accuracy and reduce the processing time [10, 11]. However, selection of Kernel functions is of importantance, as the wrong choice can lead to an tremendous increase in the error percentage. Other algorithms, such as Decision Tree, Naive Bayes Classifier, and Random Forest have lower accuracy than ANNs and SVM, but are relatively easy to use and can be used with small datasets of healthcare institutions. The Decision Tree algorithm is easy to understand and is unlikely to encounter over fitting because of the relatively small dataset. It can be used with a series of yes/no like questions to classify datasets into classifications and also be used in making of CV risk prediction in simple tasks [12]. The Random Forest algorithm is an extension of the Decision Tree algorithm, in which Decision Trees are combined and each Decision Tree is independently trained. Random Forest algorithms have been used in coronary computed tomography angiography, readmission for HF patients, and HF risk and survival prediction models [13, 14, 15, 16, 17]. In addition, random forests can easily perform leave-1-out predictions (i.e., sensitivity analysis in meta-analysis) and are relatively robust to selection bias. The Naive Bayes classifier is a simple probabilistic classifier derived from Bayes theorem. It performs very well on small training datasets and can be used in text classification problems, such as in CV risk factor identification and decision-making systems [18, 19].Fuzzy logic has been used in various places, such as for prediction of early-stage coronary artery disease, mortality prediction after cardiac surgery. [20]. K-nearest neighbor is one of the simplest nonparametric methods. It executes quickly on small training datasets and can be used in ECG interpretation problems; However, K-nearest neighbor requires more space and time for large datasets

2.3 Limitations of Unsupervised Learning

One of the major limitation of Unsupervised Learning is difficulty in identifying the initial cluster pattern, which could potentially lead to biases. Because the final cluster pattern is dependent on the initial cluster pattern, this could result in inaccuracy of decisions. Thus, it needs validation in several cohorts. In addition, some complex problems lead to limitations that are not easily solved without supervised training; thus, it may require manually labeled data to identify the optimal algorithm. Thus, for the best results, Unsupervised Learning may require manual hand coding in some of the parts, unsupervised algorithms in other parts, and subsequent validation. We expect that our proposed model will adapt to leverage dynamic treatment information in EHR data to boost the performance of prediction for ACS, and can readily meet the demand of clinical prediction of other types of diseases, from a large volume of EHR in an open-ended fashion.

2.4 Related Work

Here, we briefly introducing the background closely related to the research proposed in this project. One is clinical risk prediction from EHRs and another is the methodology research on deep learning and its applications in biomedicine.The

author suggested that the complexity of a risk prediction model increases with the increase in number of patient features and the heterogeneity of EHR data. It is important to note that the problems faced in clinical risk prediction problem with EHR data are similar to pattern classification problems that have high dimensional data. In the last decade, pattern classification has advanced into a new paradigm with emerging techniques of deep machine learning [19]. It must be noted that the models along this line have been estimated using a small set of specially chosen patient features from highly-stratified cohorts. Recently, many advanced data mining algorithms have been introduced for clinical risk prediction, with the wide spread adoption of electronic health records (EHR) [3,4,5].The EHR typically records a diverse set of clinical information, including patient demographics, symptoms, laboratory test results, and treatment behaviors, etc. It, therefore, provides a comprehensive source for clinical risk prediction. Many data mining algorithms, such as decision trees [6], Bayesian networks [7], and fuzzy inference systems [8], etc., have been proposed to explore the potential of EHR data for clinical risk prediction. For example, Tay et al. presented a novel neural inspired algorithm for risk prediction by using EHR [9] coronary heart events from EHR data [10]. In [11], a hybrid model was developed for automatically identifying risk factors of heart disease in patient. EHR.

3. PROBLEM STATEMENT

The most important task in clinical risk prediction is to develop "robust prediction models" that can effectively handle enormous heterogeneous EHR data and accurately classify various clinical risks levels based on the acquired EHR data. Deep learning architectures with a greater number of layers, can potentially extract abstract and invariant features for better performance of patient classification. The ability of inference on a large volume of heterogeneous EHR data is particularly suitable for our aim. Therefore, this paper proposes a novel approach for clinical risk prediction of ACS based on deep learning. Among various deep learning models, the Stacked Denoising Auto-encoder (SDAE) has particular advantages such as rapid infference and the ability to reconstruct features (a.k.a. clinical risk factors) yielding good classification accuracy. Thus, SDEA is adopted to address the problem of clinical risk prediction in this study. Specifically, given the complex and high-dimensional EHR data available at hand, we first developed an appropriate feature representation by training a SDAE model for patient individuals, such that a representation of ACS risk factors can be identified and the salient patient features can be disentangled from a large volume of EHR data. Our proposed model can learn more discriminate patient feature representations and thus improve the performance of clinical risk prediction.

4. PROPOSED SYSTEM

The proposed regularized SDAE keeps in memory of the characteristics of patient risk information during learning, and thus it holds the ability to enforce the reconstructed feature representations within the same risk level to be as close as possible and the reconstructed feature representations between different risk levels to be kept distant as much as possible. We apply this regularized SDAE to pre-train a clinical risk prediction model from EHR data.

The dataset is compounded by a substantial amount of EHR. Each piece of EHR, corresponding to a particular ACS patient sample, can be represented as a patient feature vector. We will discuss the training and validation of our clinical risk prediction model using our proposed approach. Typically, a SDAE model, as a symmetrical neural network, is mainly used for learning the features of a dataset in an unsupervised manner (see Figure No.1). To build a deep learning architecture with K hidden layers, SDAE is trained in a greedy layer-wise unsupervised mode. In order to build a deep learning architecture with K hidden layers, the proposed model is trained by gradient descent layer by layer, beginning with the lowest layer of SDAE. Specifically, the learning process starts by training the first DAE in an unsupervised way by optimizing with the noise input.

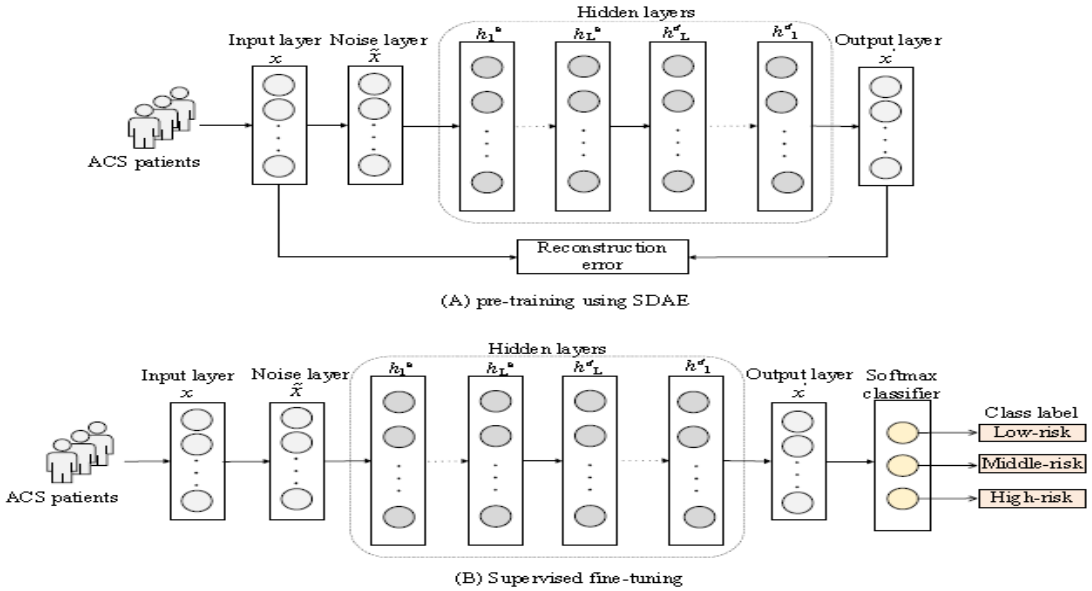

Figure No-1: System Model

Proposed system focus on a novel approach for clinical risk prediction of ACS based on Deep Learning. Among various Deep Learning models, the Stacked Denoising Auto-Encoder (SDAE) will be used. SDAE is a symmetrical Neural Network and mainly used for learning the features from data set 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. Specifically, given the complex and high-dimensional EHR data available at hand, we will first develop an appropriate feature representation by training a SDAE model for patient individuals, such that a representation of ACS risk factors can be identified and the salient patient features can be disentangled from a large volume of EHR data. In order to ensure that features reconstructed by SDAE can finally be useful to the clinical risk prediction problem, two regularization constraints that preserve actual risk information of patients are added on SDAE. One constraint penalizes a derivation of feature representations in the same risk level, to make the reconstructed features of patients within the same risk level as close as possible, and the other penalizes a derivation of representations in different risk levels, to make there constructed features of patients at the different risk levels as separated as possible. After this feature reconstruction learning phase, We will append regression layer on the top of the resulting reconstructed feature representation layer, which is tailored to the clinical risk prediction problem. The proposed regularized SDAE keeps in memory of the characteristics of patient risk information during learning, and thus it holds the ability to enforce the reconstructed feature representations within the same risk level to be as close as possible and the reconstructed feature representations between different risk levels to be kept distant as much as possible.

4.1 Methodology (RNNs)

Deep learning allows computational models which mostly composed of multiple processing layers for learning representations of data with multiple abstraction levels. This has dramatically improved machine learning performance in many domains, such as speech recognition, natural language processing, and computer vision, and has also demonstrated great performance in healthcare as well as in medical domains, such as using deep neural networks to detect referable diabetic retinopathy.

Recurrent Neural Networks, RNNs are an extension of feed-forward neural networks to model sequential data, like time series, event sequences and natural language text. In particular, the recurrent structure in RNN is able to capture the complex temporary dynamics in the longitudinal EHR data, thus making them the preferred architecture for several EHR modeling tasks, including sequential clinical event prediction, disease classification and computational phenotyping. The hidden states of the RNN work as its memory, as the current state of the hidden layer depends on the previous state of the hidden layer and the current time input. This also enables the RNN to handle variable-length sequence input. Two

prominent RNN variants with gating mechanisms are widely used: the LSTM unit, and the GRU. They are designed to overcome the vanishing gradient problem as well as to capture the effect of long-term dependencies.

The proposed model can extract informative risk factors from EHR data. Here we tried to identify the top 10 risk factors for the patient group within a particular risk level, and confirm their clinical validity with the clinical experts. Note that normalized patient features were categorized into two classes, i.e., 0 or 1 (i.e. the patient case has this feature or not), during the model learning process and then, we recovered them and show the actual values of these features, to make sure that they are understandable for clinicians. The most informative risk factors have been reviewed and endorsed by our clinical collaborators. Here we pointed out that these selected patient features are truly informative risk factors, and some of them, such as age,smoking status, etc., had already been validated within a clinical cohort study and had been recognized and adopted for ACS risk prediction. The risk factors selected by our model are in, if they are in accordance with risk factors identified previously in clinical research literature [2, 3, 9]. In addition, clinicians stated that our model can provide specific values of selected risk factors with different risk levels of ACS patients. Note that the extracted values of the same risk factor are different when they are used to indicate different risk levels. For example, the average age of high risk patients is 76.42 YO, which is higher than one of middle risk patients (69.92 YO). Thirdly, our model provides weights of these risk factors at risk levels of patients. This helps to understand the significances of risk factors given different risk levels. Moreover and most interesting, our model may find some new potential risk factors. For example, with regard to the eighth ranked informative risk factor "Abnormal liver - True" of 'high risk' patients, clinicians from the cardiology department of the hospital suggested that this might be a potential risk factor of ACS (personal communication) specific to the Chinese population. Note that Asian countries like India, China,Russia have large population with hepatitis, there could be potential correlation between ACS and Hepatitis. To the best of our knowledge, this has not previously been indicated in the literature. We will investigate this finding in their clinical study. To apply our proposed model in clinical practice, we should consider the implementation of a simpler model and investigate the compromise between simplicity and accuracy. In fact, clinicians are interested to know that how many risk factors they need to collect to obtain a good predictive performance. To this end, we apply our risk factor selection strategy to extract the most informative risk factors with regard to each risk level and feed the top-50 in the union of these risk factors directly to our proposed model for ACS risk prediction.

5. CONCLUSION

This work proposes a novel learning approach to address the clinical risk prediction problem of ACS from heterogeneous EHR data. In comparison with traditional ACS risk scoring methodologies that relied on a small set of handpicked risk factors, the proposed approach can utilize a large volume of heterogeneous EHR data to construct a robust clinical risk prediction model. The experiments were conducted on a real clinical dataset and results demonstrate that the proposed model can achieve a competitive performance in clinical risk prediction, compared with state-of-the-art classification algorithms. In addition, it has great potential to identify informative risk factors for ACS patients to different risk levels.

REFERENCES

[1] Minas A. Karaolis, et al. Assessment of the Risk Factor of Coronary Heart Events Based on Data Mining With Decision Trees, IEEE Transactions on Information Technology in Biomedicine, 2010, 14(3):559-567.

[2] Weiwei Chen, et al. Report on Cardiovascular Disease in China 201, Encyclopedia of China Publishing House, 2015, ISBN 978-7-5000-9510-1.

[3] Z. Huang, et al. A probabilistic topic model for clinical risk stratification from electronic health records. Journal of Biomedical Informatics, 2015, 58:28-36.

[4] C.J.L. Murray, A.D. Lope. Global mortality, disability, and the contribution of risk factors: global burden of disease study, Lancet, 1997,349 (9063):1436-1442.

[5] D. Mozaffarian, et al. American Heart Association Statistics Committee and Stroke Statistics – 2015 update: a report from the American Heart Association, Circulation, 2015, 131(4):e29-322.

[6] Elliott M. Antman, et al. The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. The Journal of the American Medical Association, 2000, 284(7):835-842.

[7] Charles C. Miller, et al. Risk stratification: a practical guide for clinicians. Cambridge University Press, 2001.

[8] P.M. Brindle, et al. The accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: A systematic review. Heart,2006, 92(12):1752–1759.

[9] D.C. Goff, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College ofCardiology/American Heart Association Task Force on Practice Guidelines. Circulation, 2014, 129, S49-S73.

[10] E. Boersma, et al. Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients. Circulation, 2000, 101(22):2557-2567.

[11] P.W. Wilson, et al. Prediction of coronary heart disease using risk factor categories. Circulation, 97:1837-1847, 1998.

[12] Shaun G. Goodman, et al. The expanded global registry of acute coronary events: baseline characteristics, patients with acute coronary syndromes. American Heart Journal, 2009, 158(2):193-201.

[13] M. Matheny, et al. Systematic review of cardiovascular disease risk assessment tools. Technical Report, Agency for Healthcare Research and Quality (US), 2011,9:29-54.

[14] Jessica L. Mega, et al. Rivaroxabgtan in Patients with a Recent Acute Coronary Syndrome. The New England Journal of Medicine, 2012, 366(1):9-19.

[15] S. Bandyopadhyay, et al. 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,2015, 29(4):1033-1069.

[16] S. Paredes, et al. The CardioRisk project: Improvement of cardiovascular risk assessment., 2015,9:39-44.

[17] Muxuan Liang, et al. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach, 2015, 12(4): 928-939.

[18] M.M. Al Rahhal, et al. Deep learning approach for active classification of electrocardiogram signals. Information Sciences, 2016, 345:340-354.

[19] J. Perk. European guidelines on cardiovascular disease prevention in clinical practice, European Heart Journal. 2012, 33:1635-1701.

[20] Z. Huang, et al. On mining latent treatment patterns from electronic medical records, Data Mining and Knowledge Discovery, 2015, 29(4):914–949.