Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 4267-4273*

Research Article

A Case Study on Hospital Readmission Prediction Using Deep Learning Algorithms on EHRs

B. Sushrith[a], Shashwat Kumar Dev[b], Kohinoor Jain[c], Gopichand G[d*]

[a,b,c,d]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
[d*]gopichand.g@vit.ac.in

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: In this paper, focus is made on predicting the patients who are going to be re-admitted back in the hospital before discharge using latest deep-learning algorithms is applied on the electronic health records of patients which is a time-series data. To begin with the study of the data and its analysis this project deployed the conventional supervised ML algorithms like the Logistic Regression, Naïve Bayes, Random Forest and SVM and compared their performances on different portion sizes of dataset. The final model built uses deep-learning architectures such as RNN and LSTM to improve the prediction results taking advantage of the time series data. Another feature added has been of low dimensional descriptions of medical concepts as the input to the model. Ultimately, this work tests, validates, and explains the developed system using the MIMIC-III dataset, which contains around 38000 patient's information and about 61,155 patient's data who admitted in ICU, duration of 10 years. The support from this exhaustive dataset is used to train the models that provide healthcare workers with proper information regarding their discharge and readmission in hospitals. These ML and deep learning models are used to know about the patient who is getting to be readmitted in the ICU before his discharge will help the hospital to allocate resources properly and also reduce the financial risk of patients.

In order to reduce ICU readmission that can be avoided, hospitals have to be able to recognize patients who have a higher risk of ICU readmission. Those patients can then continue to stay in the ICU so that they will not have the risk of getting admit back to the hospital. Also, the resources of hospitals that were required for avoidable readmission can be re-allocated to more critical areas in the hospital that need them. A more effective model of predicting readmission system can play an important role in helping hospitals and ICU doctors to find the patients who are going to be readmitted before discharge. To build this system here we use different ML and deep-learning algorithms. Predictive models based on huge amounts of data are made to predict the patients who are going to be admitted back in the hospital after discharge..

Keywords: Deep learning, Electronic health Records, long-short-term-memory, Readmission Prediction, Recurrent neural-networks.

1. Introduction

A patient who is admitted to the hospital, is usually monitored on the basis of the two most frequently asked questions: "What is happening now?" & "What will happen next?". The former refers to the diagnosis and monitoring of the present condition, the latter refers to the analysis and prediction of any future medical conditions. Although there are a lot of diagnostic technological devices that help to answer the first question, the technologies are less developed to answer the latter. Conventionally this is answered by experienced and highly trained physicians with well-equipped clinical data, but this method is expensive and not available to most of the populations around the world. The Electronic Health Records(EHRs) available in recent years have potential to provide fast and cheap solution. An EHRs normally contains details of patients such as time of admission, procedures, diagnoses, patient transfers and many more. Admission to Intensive Care Unit (ICU) is costly, therefore any decisions related to the timing of discharge or stepping-down to the regular ward care should ensure efficient allocation of the finite medical resources. Also, premature discharge from intensive care unit (ICU) may increase the risk of patients because of no proper monitoring and it also increase the chances of patient getting readmitted in the hospital. All of these cases together increase the chances of deterioration of patient's medical condition, and also improve the chances of death of patient and also his financial risk will be increased.

It is reported that the death rate patients who are admitting in ICU after discharge is in the range of 26% to 58%. Astonishingly the rate of patients who are admitting back to the hospitals after discharge are high even in developed countries, approximately 11% patients are admitted back to the hospitals after discharge. In the U.S. too, it has been found that ICU readmission rates have been on the rise from 4.6% in 1989 to 6.4% in 2003. All this analysis makes the rate of person re admitting to hospitals as a crucial factor in evaluating the performance.

Although various studies have worked to provide a solution in finding the patients who are at a risk and have greater probability to get admit back in the hospital but the solution was not feasible and practical. Mainly there are three disadvantages in the existing studies and proposed models.

- The present existing models are restricted in predicting the readmission for a particular disease doesn't have a generalized solution. For example, there are few models particularly for heart diseases, diabetes, HIV and kidney transplants.
- Rarely any model has the ability to predict readmissions to a reasonable level of accuracy yet; most of the models have a very less uncertainty of about 0.60 - 0.63.
- Many existing models do not completely make proper use of the time-series data which is very crucial data for building the models and also not properly utilize the various parameters of electronic health record which leads to the loss of accuracy of the model.

In recent times, deep learning is being used on various works done on electronic health records, which include correctly organized (example- Medications) as well as not properly organized (examples- clinical notes and some medical information) data about patients in the hospitals. EHR stores data accurately, decreases the risks of data copying and the risks of data loss. EHRs are being widely deployed in most developed countries and are being increasingly used in the rest of the world. Projections show that EHRs of patients in hospital would aid to improve quality of treatment and significantly reduce healthcare costs. An electronic health record includes the details of the patient and all the series of patient admitting in the hospital. Typically, there are various for patients getting admitted in the hospitals such as normal check-up type or admitted because of serious reason. The second type of patient admission consists of patients shifted from various ICU departments. EHRs normally contains details of patients such as time of admission, procedures, diagnoses, patient transfers and many more. Generally, the details such as transfers, ICU stays, test reports and prescriptions which are saved in EHRs are usually encoded in standardized formats. Generally, WHO's ICD coding schemes are used to show diagnoses.

2. Workflow and Methodology

Dataset Acquisition: MIMIC-III dataset has been used for readmission is prepared from a Clinical Database. It is one of the vast and latest database which is only accessed upon approval. It has EHR of around 40,000 patients who have been admitted in ICU from 2001 to 2012, in the medical Centre named Beth Israel Deaconess. Dataset may contain multiple records for a single patient.

Pre-Processing and cleaning of the acquired database:

The first step in the data screening process is to categorize the scheduled and unscheduled or unplanned readmissions to hospitals using the EHRs from the dataset. This is achieved by processing the admission time series column from the dataset to observe the previous hospital visits of patients. If they had a visit in the last 3 months they are categorized as high-risk patients who tend to have higher comorbidity burden. The second step involves screening those people who are below 18 years and also removing passed away people during their stay in emergency ward. This step concludes to around 36,000 people with around 48,489 ICU stays. The next step is to divide the final patient records for training -80%, testing -10% and validation -10% and dividing for model training to be able to deploy a n-fold cross-validation with n=5. The cross validation will pick a consecutive subset of patient records from the dataset which may include replacement since there are multiple records for a single patient. This will enable to enhance the training for our models.

Finally, to construct the dataset with respect to ICU readmission dataset, the next step is to categorize all these screened patients and their ICU stays records and label them as positive and negative cases. Records which are labelled as 1s are termed as positive cases. These are the records where patients get advantage of projection whether he will be readmitted or not before getting discharge.

- There are 3,768 records where people who were readmitted to ICU after getting shifted from ICU to local ward.
- There are 1,743 records where people died after getting shifted from ICU to local wards.
- There are 2,089 records where people readmitted to ICU in less than 30 days after getting discharge.
- There are 2,643 records where the people died in less than 30 days after getting discharge.

Records which are labelled as 0s are termed as Negative cases. These are the cases where ICU readmission is not important for the patients. In general, we can say these are the people that were released and not readmitted but lively for more than a month.

Feature Extraction

At this stage it is required to choose the suitable data to be used for the prediction task that is for ICU readmissions. With respect to the temporal data modelling of the sequence of numerical data points records of each patient staying for last 2 days. This is because the data recorded during last 2 days until the release of patient is proved best way for data prediction.

The three categories used in this project are:

• **Electronic Chart Events** are the events from the notes of the health professionals like doctors, dentists, nurses etc. This category shows the patients' medical report on the basis of laboratory measurements. The major chunk of 17 kinds of sequence of points are extracted from this category of events for a 48-hour period for all patients in the selected category.

• **ICD-9 embedding's,** represent each patient's information factors like infections or diseases. This observation is been discovered to be highly linked with the chances of readmission. In dataset of EHR's the data was appeared to be scattered, thereby seems to be difficult task to do it with deep learning.

• **Demographic information** which includes people's age, gender and income. These categories have also been considered as crucial factors in the prediction of readmission. The insurance type is also included in this data as there may be cases when an uninsured person is discharged from the hospital due to insufficient payment.

Statistical features for conventional ML models

In order to be able to perform analysis using the conventional machine learning algorithms, it is required to calculate the statistical characteristics of 2-day period. The a and b values on the regression line which indicates slope and intercept are evaluated as discrete characteristics for distinguishing between linear data, since this strategy is previously majorly used for the same purpose. The mean and modal values throughout the all-out time window are determined and utilized after changing labelled events into binary or ordinal to assess straight out information.

3. Machine Learning Models Architecture

Baseline Conventional Models

Logistic Regression: The first baseline model that the data is trained on is the logistic regression model. The execution is done on the multiple variants of logistic regression including L1 and L2 (Lasso and Ridge) which is an enhancement to the regular version for preventing overfitting and high covariance in the variables of the dataset. In generally to label the data we use logistic regression. for bringing down value of coefficients in the outcome, lasso and ridge methods are used.

Since the variables which are used as input are manipulated, therefore there has been a decreasing in variance.

Naïve Bayes:

It is a commonly used model which uses the concept of bayes theorem for prediction. For our model training after the feature selection the document-term sparse matrix was created by vectorization and using the bag of words technique popular to this classifier. It helps to train this model better.

Random Forest: This is also a supervised ML classifier that is an addition to the decision tree based approach which works on Information Gain, Entropy and Gain Index. The problem with that approach was that it led to overfitting which is solved by Random Forests, this algorithm is a cluster of trees which generates results on the vote of each tree in the forest. Less number of trees will not produce good results. Therefore, to create a greater number of trees even from a small dataset it randomly selects the chunks of data with replacement and assigns to each decision tree. For predicting an unknown class for a data value, the class predicted by most number of trees is selected as the predicted value.

Support Vector Machines (SVM): This ML model is basically used for linear data but works well for non-linear data with some intelligent modifications. It basically creates a decision boundary for a bi-class multi-dimensional data. For our data the values calculated by statistical evaluation on time-series features is used as input which are of numerous dimensions. Hence, to map and train the model effectively so as to keep the margin as big as possible the gamma, C and kernel values are chosen appropriately. Kernel value is set to 'rbf' to allow this model to decide what's the best loss function for the training values.

Deep Learning Based Models

Convolutional neural network (CNN) model:

To distinguish from other model, a CNN-based model is also implemented to differentiate from the other models. CNN-based approaches are studied and proved to be beneficial in analyzing EHR time series data. The training of the model is done on both comprehensive and longitudinal depiction of data with 19,715 selected dimensions. The convolution is applied with respect to time axis with 48-hour time period and D dimension employing filter sizes of two, three or four accordingly.

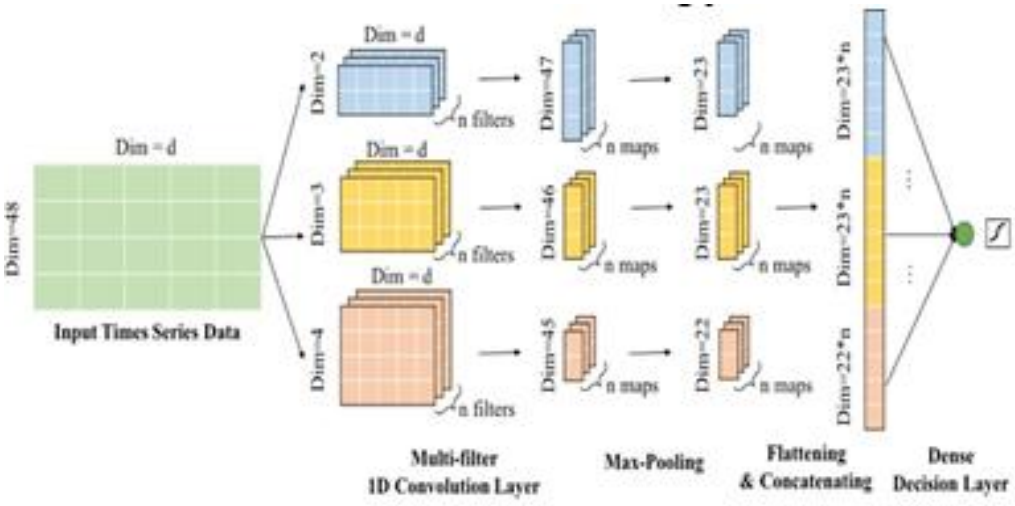

Figure 1: CNN Model Used in this project for prediction

Recurrent Neural Network:

Although this project does not directly deploy the regular version of this artificial neural network, a latest and much more refined variant of RNNs known as the LSTM-RNN is used for the project, due to the capabilities of handling temporal data much more efficiently and overcoming all the drawbacks of the original, raw version. RNNs prove to be excellent for solving problems where the order of the data matters. In our case the final 48-hour period of the patient's ICU stay is order specific and also time series based.

Long short-term memory (LSTM) model:

LSTM networks is observed adequate in prediction forming built over time series data, particularly for clinical measurements in which there is delay possibility due to unfamiliar duration and missing values in a time series. A combination of three layers bidirectional LSTM, LSTM and a dense decision layer with a single resulted neuron is used. Overall, there are sixteen concealed units in the constructed LSTM layer.

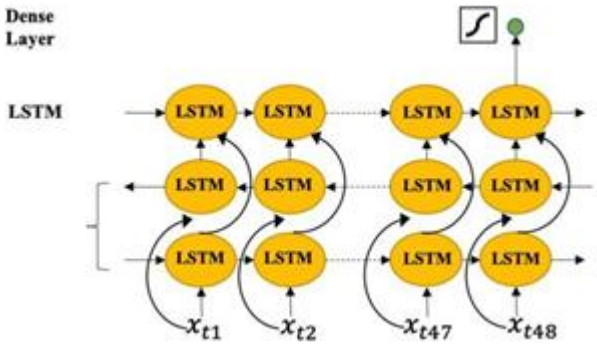

Figure 2: Bi-directional LSTM network representation

4. Result And Discussion

This segment, illustrates all the trials and processes that have been executed to examine how the predictive models performs.

• The first part includes the results for all the conventional models (Naïve Bayes, random forest, LR, SVM), in addition with the temporal LSTM models and deep learning techniques of convolution neural networks applied on the demo MIMIC III data obtained from the Physio Net portal.

• After successfully testing the working of these models and the logic tweaking done in the codes for these, the models are used to predict the whole dataset which is about 6.5 GB in size when uncompressed, and therefore takes a lot of computational resources and significant amount of time to train the predictive models.

The results for these are summarized in the later part of this section. Finally, a comparison is made among all the obtained results and the most optimal ICU readmission prediction solution is selected.

Table 1: Summary of Results of Regression Based Models

Model	Features Combination	Recall (95% confidence)	AUC-ROC (95% confidence)
LR-L2	L48-hours + ICD9	0.776 (0.759–0.793)	0.875 (0.861–0.894)
LR-L2	L48-hours + ICD9 + Dem	0.785 (0.761–0.810)	0.884 (0.873–0.895)
LR-L1	L48-hours+ ICD9	0.778 (0.765–0.792)	0.886 (0.876–0.897)
LR-L1	L48-hours + ICD9 + Dem	0.791 (0.773–0.808)	0.888 (0.876–0.900)
Model	Features Combination	Recall (95% confidence)	AUC-ROC (95% confidence)
NB	L48-hours + ICD9	0.453 (0.434–0.472)	0.709 (0.702–0.716)
NB	L48-hours + ICD9 + Dem	0.509 (0.479–0.540)	0.706 (0.698–0.713)
RF	L48-hours + ICD9	0.563 (0.548–0.578)	0.714 (0.703–0.725)
RF	L48-hours + ICD9 + Dem	0.565 (0.550–0.580)	0.712 (0.693–0.730)
SVM	L48-hours + ICD9	0.701 (0.686–0.715)	0.786 (0.776–0.796)
SVM	L48-hours + ICD9 + Dem	0.703 (0.685–0.720)	0.784 (0.779–0.790)

Table 2: Results from Baseline Classification Models

Model	Features Combination	Recall (95% confidence)	AUC-ROC 95% confidence
CNN	L48-hours + ICD9	0.665 (0.586–0.745)	0.780 (0.774–0.786)
CNN	L48-hours + ICD9 + Dem	0.735 (0.676–0.794)	0.784 (0.773–0.794)
CNN+LSTM	L48-hours + ICD9	0.739 (0.670–0.807)	0.785 (0.775–0.795)

Table 3: Results from LSTM model with different combinations of features

Model	Features Combination	Recall (95% confidence)	AUC-ROC 95% confidence
LSTM	L48-hours + ICD9	0.717 (0.692-0.742)	0.784 (0.772–0.795)
LSTM	L48-hours	0.593 (0.537–0.649)	0.704 (0.697–0.710)
LSTM	L48-hours + ICD9+Dem	0.733 (0.698–0.768)	0.787 (0.771–0.802)

Table 4: Results from combination of LSTM and CNN models

Model	Features Combination	Recall (95% confidence)	AUC-ROC (95% confidence)
CNN	L48-hours + ICD9	0.665 (0.586–0.745)	0.780 (0.774–0.786)
CNN	L48-hours + ICD9 + Dem	0.735 (0.676–0.794)	0.784 (0.773–0.794)
CNN+ LSTM	L48-hours + ICD9	0.739 (0.670–0.807)	0.785 (0.775–0.795)
CNN+ LSTM	L48-hours + ICD9 + Dem	0.710 (0.648–0.771)	0.787 (0.775–0.799)
LSTM +CNN	L48-hours + ICD9	0.729 (0.647–0.811)	0.786 (0.776–0.796)
LSTM +CNN	L48-hours + ICD9 + Dem	0.742 (0.718–0.766)	0.791 (0.782–0.800)

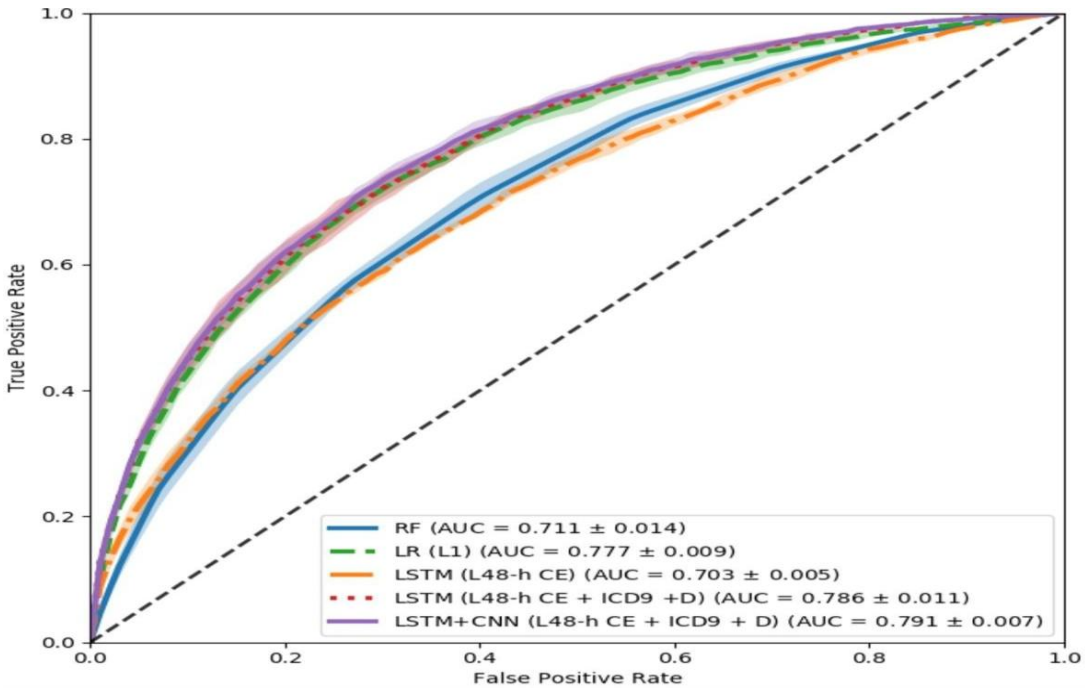

Figure 3: Comparison of Final ROC Curves for different models. error bar of the is indicated by colored bar

5. Conclusion

In this paper, demonstration of the unplanned hospital readmissions prediction by making the use of electronic health records, ICD-9 embedding's along with demographic features provided in the MIMIC dataset has been shown. Amongst all the types of data that has been used, it is observed that features of the EHRs are suggestively responsive to time series, and therefore cannot be depicted accurately through traditional ML models (e.g. Naïve Bayes, logistic regression). Therefore, combination of LSTM (Long short-term memory) and CNN models is deployed that seems to efficiently include temporal data in the absence of which much information is lost. This deep learning tool that predicts hospital readmissions possibly produces greater sensitivity and accuracy in

contrast to other predictive models. Additionally, this model can be deployed at different operating values it can be modified to match the required sensitivity and specificity, for usage in critical care. The final values for AUC and sensitivity achieved by this model are 0.79 and 0.742 respectively. Further, the importance and weightage corresponding to every input feature along with its different combinations affecting the results of the predictive model have been tried and demonstrated.

References

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, Joel T Dudley, Deep learning for healthcare: review, opportunities and challenges, Briefings in Bioinformatics, Volume 19, Issue 6, November 2018.

Miotto R, Li L, Kidd BA, et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6:26094.

Che Z, Kale D, Li W, et al. Deep computational phenotyping. In ACM International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 2015, 507–16.

Liang Z, Zhang G, Huang JX, et al. Deep learning for healthcare decision making with EMRs. In IEEE International Conference on Bioinformatics and Biomedicine, 2014, 556–9.

Tran T, Nguyen TD, Phung D, et al. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). J Biomed Inform 2015;54:96–105

Nguyen P, Tran T, Wickramasinghe N, et al. Deepr: a Convolutional Net for Medical Records. IEEE J Biomed Health Inform 2017;21:22–30.

Choi Y, Chiu CY-I, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. AMIA Jt Summits Transl Sci Proc. 2016; 2016: 41–50. PMID: 27570647

Miotto R, Li L, Dudley JT. Deep learning to predict patient future diseases from the electronic health records. In European Conference in Information Retrieval, 2016, 768– 74.

Razavian N, Marcus J, Sontag D. Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests [Internet]. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1608.00647

Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One 2013;8:e66341.

Pham T, Tran T, Phung D, et al. DeepCare: a deep dynamic memory model for predictive medicine. arXiv 2016. https:// arxiv.org/abs/1602.00357.

Andrea Gruneir, Irfan A D, Carl van W, et al. Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm, Sydney, NSW, Australia, 2011, 104–11.

Lin Y-W, Zhou Y, Faghri F, Shaw MJ, Campbell RH (2019) Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLoS ONE 14(7): e0218942.

Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ, et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a crosssectional machine learning approach. BMJ Open. 2017; 7: e017199. https://doi.org/10.1136/bmjopen-2017-017199 PMID: 28918412