A deep learning approach for length of stay prediction in clinical settings from medical records

Tahmina Zebin
School of Computing Sciences
University of East Anglia
Norwich, UK
Email: t.zebin@uea.ac.uk

Shahadate Rezvy
School of Science & Technology
Middlesex University London, UK
Email: s.rezvy@mdx.ac.uk

Thierry J. Chaussalet
School of Computer Science and Engineering
University of Westminster
London, UK
Email: chausst@westminster.ac.uk

***Abstract*—Deep neural networks are becoming an increasingly popular solution for predictive modeling using electronic health records because of their capability of learning complex patterns and behaviors from large volumes of patient records. In this paper, we have applied an autoencoded deep neural network algorithm aimed at identifying short(0-7 days) and long stays (>7 days) in hospital based on patient admission records, demographics, diagnosis codes and chart events. We validated our approach using the de-identified MIMIC-III dataset. This proposed Autoencoder+DNN model shows that the two classes are separable with 73.2% accuracy based upon ICD-9 and demographics features. Once vital chart events data such as body temperature, blood pressure, heart rate information available after 24 hour of admission is added to the model, the classification accuracy is increased up to 77.7%. Our results showed a better performance when compared to a baseline random forest model.**

***Keywords*-Deep learning; Electronic Health Records; Clinical Prediction; Length of Stay.**

I. Introduction

Health care observations, stored in electronic health records (EHR) are episodic and irregular in time [1], [2]. Additionally, the evolving nature of clinical practice in health care systems results in clinical prediction models being outdated and less accurate over time. A recent review of predictive models built with EHR data reported that these are mostly linear models that are not flexible to changes in features [3]. These traditional predictive modeling techniques require the creation of a custom dataset and handcrafting of features to extract and normalize. However, as patient records vary significantly in length and density with various medical parameters being recorded over a period of time in EHR databases, statistical and linear models become quite limited in terms of scalability [4]. A potential alternative is to develop prediction models utilizing deep neural network architectures to facilitate automatic feature extraction in a data driven manner. These models can retain accuracy by evolving over time in response to observed changes. In this paper, we create a Dense Neural Network(DNN) model that predicts the length-of-stay (LOS) for each patient at the time of admission.

In recent literature, deep learning techniques have been successfully implemented for diagnoses[5], [6], mortality prediction and estimating length of stay from EHR[4], [7]. Amongst these, LOS is an important metric for assessing the quality of care and planning capacity within a hospital. The length of time patients spend in hospital is a good representation of their resource usage (e.g. bed capacity, equipment and staff requirement) [8]. Effective use of resources is of immense importance to health care decision makers nowadays [9]. The input data used to predict LOS usually includes a combination of basic information about the person such as age and gender, medical data, lab results, demographic data and other circumstances regarding the admission. Some typical examples of demographic data that might be used are race, expected primary payer (type of insurance if there is one), living arrangement and marital status. Other circumstances regarding the admission can be mode of arrival (walk-in, ambulance) and urgency of surgery. Medical data can include diagnoses, physical symptoms and injuries, drug records, number of drugs taken daily, other diagnoses than the cause of hospitalization, surgical history and other medical history, family history of some diseases, images such as X-rays etc [10], [11].

In this work, we utilize the Medical Information Mart for Intensive Care III (MIMIC-III) dataset as the benchmark EHR. Our Length of stay classification scheme is based upon an autoencoded dense neural networks to obtain better performance compared to using a dense neural network. The rest of the paper is organized as follows. Section II introduces a literature review on the length of stay problem and MIMIC III data selection procedure for the classification scenario of this paper. Section III presents our proposed model architecture, with results presented and discussed in Section IV. Finally conclusions are drawn in Section V.

II. Background

A. MIMIC-III Overview

MIMIC-III is a publicly available database that comprises Electronic Health Record (EHR) information related to almost 40000 patients admitted to critical care units at the Beth

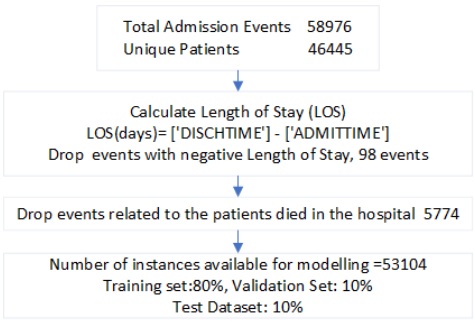

Figure 1. Inclusion criteria for events in length of stay modelling dataset

Israel Deaconess Medical Centre, in Boston, USA between 2001 and 2012 [12]. It consists of 26 relational tables, where 16 of them contain time stamped event information. It contains data such as: vital signs, medications, laboratory measurements from within the hospital (i.e. in-patient) and from clinics (i.e. out-patient), charted observations during a patients stay in the intensive care unit, and de-identified notes regarding the patients stay, including nurse notes, physician notes and discharge summaries. Tables are linked by identifiers: $SUBJECT\_ID$ refers to a unique patient and $HADM\_ID$ refers to a unique admission. For this study, we have mainly used chartevents table, along with linked d_item table to get the label of $item_id$ specified in the chartevents table. Diseases and procedures in the MIMIC-III dataset are encoded using the International Classification of Diseases version 9 (ICD-9) codes, and the mapping can be found in $diagnoses\_icd$ and $procedures\_icd$ tables. All the patient data in the MIMIC-III database has been de-identified and all dates have been randomly shifted to the future so that dates are internally consistent for the same patient but inconsistent across patients [13].

B. *Features and output classes*

In order to predict hospital LOS, the MIMIC data needed to be separated into a dependent target variable (length-of-stay in this case) and independent variables (features) to be used as inputs to the model. The data needed significant cleanup and feature engineering to be in a format compatible with the learning model. Pandas and scikit-learn libraries for Python were used for cleaning up the data. We excluded erroneous data such as negative length of stay instances, which may have been caused by data entry error in the admit and discharge times. We have also excluded events related to the patients dying in the hospital. The criteria for an event to be included in the LOS modelling task are summarized in Fig. 3

In the next subsection, we provide brief details of features and target classes used for this length of stay modelling task.

1) Admission info, Demographics, ICD-9, and Chart Events: The first group of variables are patient variables, demographics and diagnosis code related diseases, especially chronic diseases, that are found strongly associated with LOS in hospital. The demographic features we included from the MIMIC III dataset are gender, age, race, marital status and insurance type. Basic demographic information, such as gender, age, and race are usually shown as important factors in state of the art LoS prediction. Our reason for including insurance type (uninsured) was that being uninsured could lead to insufficient payment and result in premature discharge. All these variables are categorically encoded for the modelling purposes.

Chronic diseases are one of the most important factors associated with longer length of stays(>7 days). To deal with the EHR ICD-9 codes, as it was not feasible to have all the 6984 unique values to use as features for predicting LOS, we combined diagnoses into the following seventeen super categories: infectious and parasitic diseases (codes 100 - 139); neoplasms (140 - 239); endocrine, nutritional and metabolic diseases, and immunity disorders (240 - 279); diseases of the blood and blood-forming organs(280 - 289); mental disorders (290 - 319); diseases of the nervous system and sense organs (320 - 389); diseases of the circulatory system (390 - 459); diseases of the respiratory system (460 - 519); : diseases of the digestive system ; diseases of the genitourinary system (580 - 629); complications of pregnancy, childbirth, and the puerperium (630 - 679) ; diseases of the skin and subcutaneous tissue (680 - 709); diseases of the musculoskeletal system and connective tissue (710 - 739); congenital anomalies (740 - 759); certain conditions originating in the perinatal period (760 - 779); symptoms, signs, and ill-defined conditions (780 - 799) , injury and poisoning (800 - 999) and miscellaneous disease category etc. The median LOS for various disease groups in the dataset is shown in Fig. 2 (b).

In this LOS prediction study, we considered the 24 hour after admission chart events information along with the demographic and ICD-9 features. We included seven chart events (e.g. weight, height, pH, respiratory rate, body temperature, systolic and diastolic blood pressure) to attempt to improve the performance of the models in predicting whether a patient stay will be long or short.

2) Creating the target classes: The threshold value for the class division was chosen to be seven, as the LOS histogram presented in Fig. 2(a) shows a median LOS of 6.56 days. The entire dataset was labelled into two broad groups of short stay (0-7 days) and long stay (>7 day). We have also conducted various exploratory analyses of features to gain insight into variations in median LOS across categories. These variations are visible across the seventeen diagnosis super categories on Fig. 2(b), with skin and infectious super categories yielding the highest median LOS. Patients were randomly split into training (80%), validation (10%), and test (10%) sets. On the test set we have 2768 instances for short stay (0-7 days), 2568 instances for long stays (>7 day).

A feature importance analysis was conducted using the

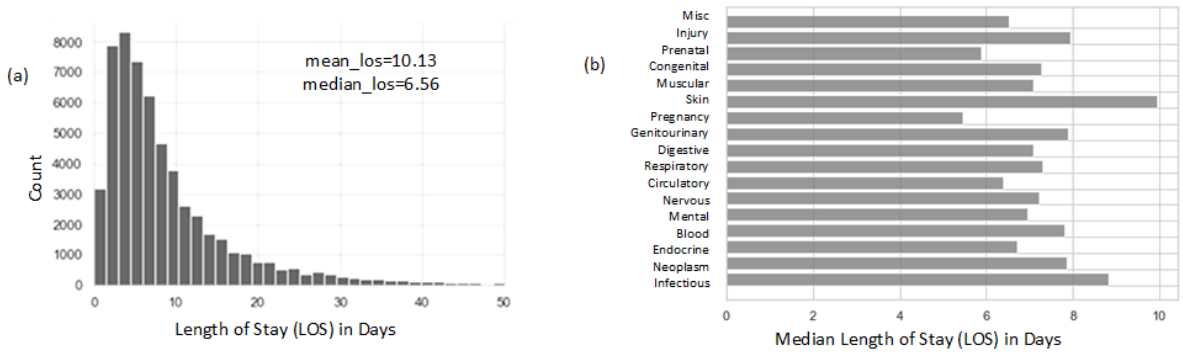

Figure 2. Exploratory analysis of length of stay (a) length of stay distribution (b) Median length of stay for various ICD -9 disease groups

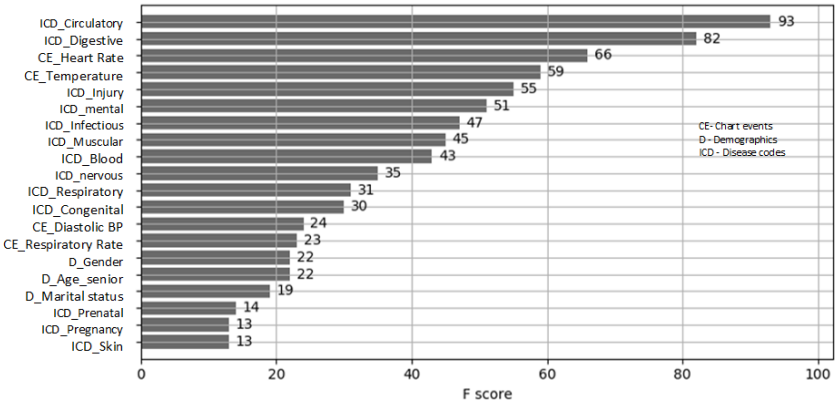

Figure 3. Top twenty features from the ICD-9, Demographics, and 24 hour Chart Events feature set

'XGBRegressor' from the python xgboost library. The plot in Fig. 3 shows the top 20 features according to their importance (expressed as F score) in model discrimination from the ICD-9, Demographics and 24 hour Chart Events feature set.

III. Model Description

To predict the admission events from the features described from the previous section, we propose a classification scheme based upon an autoencoded dense neural networks. We have implemented the model using python keras library [14] with TensorFlow back end. All of our evaluations were performed on a Linux machine with an Intel Xeon 3.60GHz processor, 128 GB RAM and an NVIDIA Titan V GPU.

1) Unsupervised pre-training with Autoencoder: An autoencoder is a type of artificial neural network used to learn efficient data representation in an unsupervised manner. In our proposed model, we have employed an autoencoder with an encoding and a decoding layer that has been trained to minimize the reconstruction error. This incorporated prior knowledge from the training set to effectively learn from the data itself and provided good performance. Such pre-training allows both the data for the current task and for previous related tasks to self-organize the learning system and build it in a data driven fashion. We have fed the autoencoder with the features from the training dataset without labels (unsupervised). A set of compressed and robust features is built at the end of this step. The encoder part of the autoencoder aims to compress input data into a low-dimensional representation, while the decoder part reconstructs input data based on the low-dimension representation generated by the encoder.

2) Supervised Classification with DNN: After the autoencoder layer, a three layer dense neural network is trained by using the outputs of the first autoencoder as inputs. This task sequence is retrained in a supervised manner with the class labels and the input feature given to the classifier. We have used a softmax activation layer as the output layer. The layer calculates the loss between the predicted values and the true values, and the weights in the network are adjusted according to the loss.

The simple *softmax* layer, which is placed at the final layer, can be defined as follows:

$$P\left(c|x\right) = \text{argmax}_{c\in C}\frac{\exp(x_{L-1}W_L + b_L)}{\sum_{k=1}^{N_C}\exp(x_{L-1}W_k)}, \qquad (1)$$

where c is the number of classes, L is the last layer index, and N_C is the total number of class types including short and long LOS. In (1), x, W and b indicates input, weight and bias associated to the final layer respectively. The learning rate of training was set to 2×10^{-3}, and we used Adam optimizer to train the final version of the model (beta = 0.9). After this stage, all layers are fine-tuned through back-propagation in a

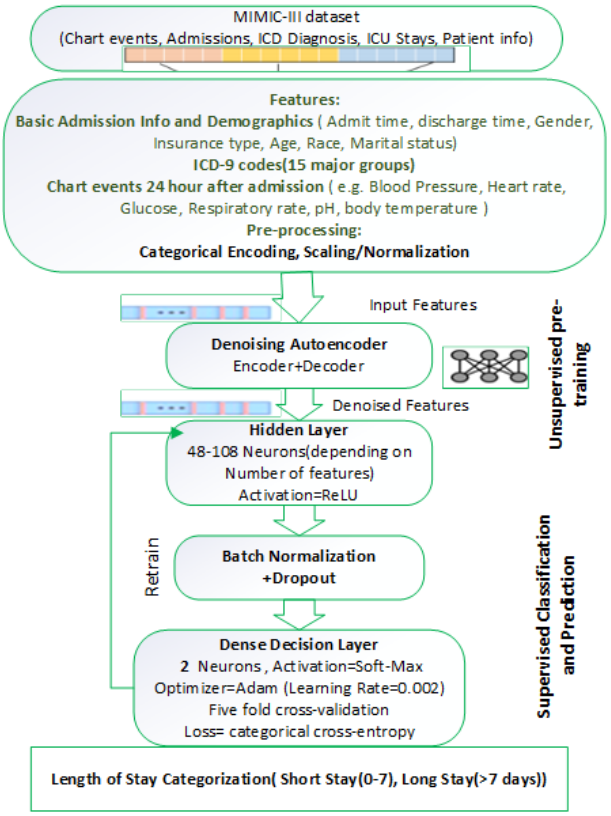

Figure 4. Autoencoder+DNN model, and a denoising autoencoder computes learns from the dtata in an unsupervised manner and the final classification was obtained using a supervised dense neural network.

supervised way. In the test phase, the softmax layer outputs the probability of the predicted categories.

IV. Results and Discussion

A. Performance Evaluation

To illustrate the two groupings in the length of stay, we present a comparative confusion matrix plot in Fig. 5, where the model is evaluated with the test data set. The diagonal cells in the confusion matrix correspond to observations that are correctly classified (True Positive: T_P and True Negative: T_N). In our case, the T_P and T_N's indicate the number of stays classified correctly as short stays (0-7 days) and long stays (>7 days) respectively. Similarly false Positives (F_P) and False Negatives (F_N) represent false detection from each group. From the confusion matrix, the overall accuracy, precision and recall performance of the model can be calculated using the following equations:

$$\text{Recall or True positive rate} = \frac{T_P}{T_P + F_N}. \tag{2}$$

$$\text{Precision} = \frac{T_P}{T_P + F_P}. \tag{3}$$

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}. \tag{4}$$

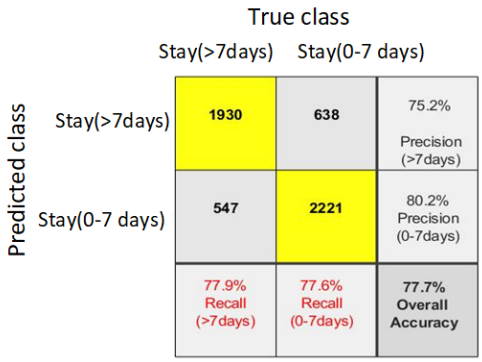

Figure 5. Class-wise confusion matrix of the test dataset

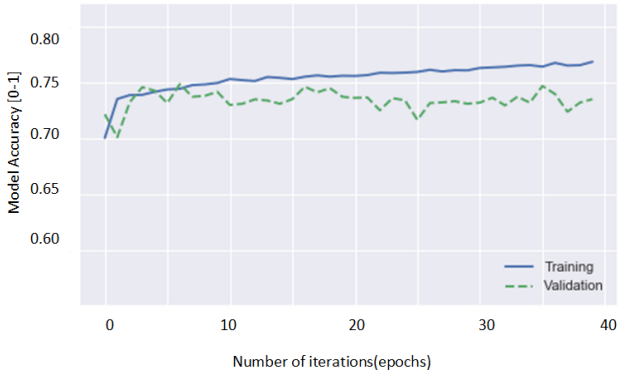

Figure 6. Training and validation set accuracy over increasing number of epochs(training iterations)

In Fig. 5, the off-diagonal cells correspond to incorrectly classified observations (F_P and F_N's). The number of observations are shown in each cell, and the overall accuracy is presented in the right lower box.

The proposed model with the auto-encoder along with a dense neural network technique identified the two classes with 73.2% accuracy based upon ICD-9 and demographics features alone. Once vital chart events information (body temperature, blood pressure, heart rate information) that is observed at admission is added to the model, the classification accuracy increases to 77.7%. Compared to a simplistic dense neural network, adding an unsupervised autoencoder at the first stage of the model substantially improved the classification performance.

During training, we have used additional techniques such as dropout, batch normalization process in place to avoid over fitting and also to speedup the training process. The proposed algorithm achieves approximately 77% accuracy for the training set in 40 iterations (shown in Fig. 6). We used 10% of the training data for retraining and validation purposes. Potentially the proposed model allows a reduction in the training epochs required to achieve the required performance. This speeding up will be of vital importance for developing low-latency models and training future networks with bigger data sets such as the Hospital Episode Statistics [15] and other linked datasets.

Table I
QUANTITATIVE COMPARISON OF AUTOENCODER+DNN WITH OTHER TRADITIONAL CLASSIFICATIONS. ON THE TABLE D: DEMOGRAPHIC FEATURES, ICD-9 : DIAGNOSES CODE FEATURES (15) AND 24H-CE: 24 HOUR LOG OF CHART EVENTS

Classifier	Features	Accuracy	Precision(0-7 days)	Precision(>7 days)	Recall(0-7 days)	Recall(>7 days)
Random Forest	ICD-9+D	66.2%	70.1%	67%	74.2%	33%
DNN	ICD-9 + D	70.2%	72.8%	62.4%	70.4%	61.8%
Autoencoder+DNN	ICD-9 + D	73.2%	73.8%	67.4%	72.4%	71.8%
Autoencoder+DNN	ICD-9 + D +24-h CE	77.7%	80.2%	75.2%	77.9%	77.6%

B. Comparison with baseline statistical models

To compare the proposed model with a baseline traditional model, a random forest classifier was trained with the same feature sets. Statistical analyses and baseline models were done in Scikit-learn Python. The baseline model was trained using the basic ICD-9 and demographics feature set to assess the effect of the variables. A comparative summary of the all the implemented models is presented in Table I. The random forest model recall accuracy (33%) in predicting longer stays was worse than any of the DNN and Autoender+DNN based implementations.

V. CONCLUSION

In this work, we applied a data-driven Auto-encoder along with a dense neural network technique aimed at identifying patients short and long stays in hospital based on their admission records, demographics, diagnosis codes (ICD-9) and chart events after 24 hour of admission data from the MIMIC III dataset. This proposed Autoencoder+DNN model showed that the two classes are separable with 77.7% accuracy based upon ICD-9 and demographics features and vital chart events data such as body temperature, blood pressure, heart rate information available after 24 hour of admission. Adding an unsupervised autoencoder at the first stage of the model substantially improved the performance compared to using a simplistic dense neural network for the classification task. The results presented in this paper could be used with other existing tools to provide complementary information to clinicians in the management of data-driven decision-making. In the future, we would like to improve our model to adapt to the realistic context of high data imbalance. We will also work on heterogeneous ensemble model that can combine multiple predictions to make more accurate predictions in clinical settings.

ACKNOWLEDGEMENT

We would like to acknowledge the support of Quintin Hogg Trust for funding this research and NVIDIA Corporation with the donation of the Titan V GPU.

REFERENCES

[1] T. Pham, T. Tran, D. Phung, *et al.*, "Predicting healthcare trajectories from medical records: A deep learning approach," *Journal of Biomedical Informatics*, vol. 69, pp. 218–229, 2017.

[2] D. A. Jenkins, M. Sperrin, G. P. Martin, *et al.*, "Dynamic models to predict health outcomes: Current status and methodological challenges," *Diagnostic and Prognostic Research*, vol. 2, no. 1, p. 23, Dec. 2018.

[3] M. J. Pencina, B. A. Goldstein, A. M. Navar, *et al.*, "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 198–208, May 2016.

[4] A. Rajkomar, E. Oren, K. Chen, *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[5] Z. C. Lipton, D. C. Kale, C. Elkan, *et al.*, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[6] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.

[7] H. Harutyunyan, H. Khachatrian, D. C. Kale, *et al.*, "Multitask learning and benchmarking with clinical time series data," *arXiv preprint arXiv:1703.07771*, 2017.

[8] E. M. Carter and H. W. Potts, "Predicting length of stay from an electronic patient record system: A primary total knee replacement example," *BMC medical informatics and decision making*, vol. 14, no. 1, p. 26, 2014.

[9] S. Curto, J. P. Carvalho, C. Salgado, *et al.*, "Predicting icu readmissions based on bedside medical text notes," in *2016 IEEE International Conference on Fuzzy Systems*, IEEE, 2016, pp. 2144–2151.

[10] C. Gholipour, F. Rahim, A. Fakhree, *et al.*, "Using an artificial neural networks (anns) model for prediction of intensive care unit (icu) outcome and length of stay at hospital in traumatic patients," *Journal of clinical and diagnostic research: JCDR*, vol. 9, no. 4, OC19, 2015.

[11] P.-F. J. Tsai, P.-C. Chen, Y.-Y. Chen, *et al.*, "Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network," *Journal of healthcare engineering*, 2016.

[12] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160 035, 2016.

[13] T. Desautels, R. Das, J. Calvert, *et al.*, "Prediction of early unplanned intensive care unit readmission in a uk tertiary care hospital: A cross-sectional machine learning approach," *BMJ Open*, vol. 7, no. 9, 2017.

[14] F. Chollet. (2013). Keras: The python deep learning library, [Online]. Available: https://keras.io/.

[15] NHS. (2019). Hospital episode statistics, [Online]. Available: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics.