Interpretable Deep Learning Framework for Predicting all-cause 30-day ICU Readmissions

Parvez Rafi
rafiparvez@tamu.edu

Arash Pakbin
a.pakbin@tamu.edu

Shiva Kumar Pentyala
shivakumar.pentyala@tamu.edu

***Abstract*—ICU readmissions are costly and most of the early ICU readmissions in the United States are potentially avoidable. After the US Govts push towards reducing avoidable readmissions, there has been a surge in research and analyses for reducing the readmission rates. Widespread adoption of Electronic Health Records(EHRs) has made large amount of clinical data available for analysis. It has provided new opportunities to discover meaningful data-driven characteristics and implement machine learning algorithms. Sequential characteristics present in EHR data can be harnessed using state-of-the-art deep learning algorithms. While there has been rapid adoption of deep models in many domains, in Healthcare sector however, their adoption has been slow owing to lack of interpretability of these black-box models. Hence, many clinical applications still prefer simple but interpretable machine learning models. In this project, we have implemented a Knowledge-Distillation approach called Interpretable Mimic Learning for predicting 30-day ICU readmissions. Using this approach, the knowledge of deep models can be transferred to simple and interpretable models and we can combine accuracy and sequential learning of deep models with interpretability of simple models.**

Keywords*—*ICU Readmissions, Deep Learning model, Interpretability, RNN, LSTM.

I. Introduction

An Intensive Care Unit (ICU) is a special unit in hospitals where people with severe and life-threatening illnesses and injuries are transferred to be provided with critical and intensive treatment. After specific criteria are met, the patients are discharged from ICUs and returned to their wards. Intensive care is expensive[2, 8, 18, 20], and accounts for around 30% of total hospital costs and 1% of the US gross national product [18, 20]. This calls for prudent decision making regarding discharging patients from ICUs. If patients are discharged prematurely, it may result in inadequate levels of monitoring as well as early readmissions to ICUs [8, 7, 19, 14]. Therefore, predicting 30-day ICU readmission of patients would not only strengthen clinical decision making about whether a patient should be discharged from the ICU, but can also save significant amount of treatment costs. Widespread adoption of Electronic Health Records (EHR) by the hospitals in over last 10 years has provided a great opportunity to develop clinical decision support systems by analyzing digital data of patient vitals, lab results, demographics and past diagnoses.

EHR data can be represented as temporal sequences of high-dimensional clinical variables where the sequence ensemble represents the documented content of medical visits from a single patient. When dealing with sequential features, simple machine learning models like logistic regression and decision trees summarize them into aggregate features, ignoring the temporal and sequential relationships among the feature elements. Although recurrent neural networks (RNN) have been successfully applied in modeling sequential EHR data[16] to predict diagnoses and model encounter sequences[9, 5] but the outputs of such models are difficult to interpret. Following the recent progress in deep learning, researchers and practitioners of machine learning are recognizing the importance of understanding and interpreting what goes on inside these black box models. Many attempts have been made to directly interpret the outcome of recurrent neural networks in the fields of speech recognition and translation, and these powerful models are also found to be very useful in the applications involving sequential data [13]. However, adoption has been slow in applications such as health care, where practitioners are reluctant to let an opaque system make crucial clinical decisions. Considering the power of RNNs for analyzing sequential data/long-term temporal properties and to overcome the trade-off between interpretability and accuracy, we have implemented a *Knowledge-Distillation Approach* [11] for predicting 30-day ICU readmissions, which preserves RNNs accuracy and sequential modelling ability while allowing a higher degree of interpretation. This is achieved by using a student-teacher learning model called *Interpretable Mimic Learning* proposed by [4] in which a slow but accurate deep model(teacher model) learns complex and sequential features and transfers this knowledge to a fast and interpretable model. This is achieved by using soft prediction scores of the deep/teacher model as target labels while training the student/interpretable model. After multiple experiments and training our models using this approach, we generate interpretable results of predicting ICU readmissions, which may help clinicians in making better clinical decisions and thus reducing the occurrence of early readmissions.

II. Related Work

Our work is built on top of a rich body of previous work on interpretable machine learning models. Choi et al. (2016) [6] stated that traditionally choice is made in between accuracy of complex black-box models such as recurrent neural networks (RNN) and interpretability of less accurate traditional models such as logistic regression. Papers have been published stating this trade-off between accuracy and interpretability [12, 3]. The most popular notions of interpretability hinge upon the intelligibility of the features (Lipton, 2016)[15]. Simple models like decision trees and logistic regression produce results that are interpretable [21] by humans but they ignore the temporal relation among features. Therefore the model accuracy is not sufficiently high. Complex models like RNNs which are known to give good results for sequential data often have limited interpretation due to their complex structure. This tradeoff poses challenges in Health care where both accuracy and interpretability are important. Interpretability is important to develop trust upon decisions[17]. To alleviate this tradeoff, Che et al. (2016) [4] proposed a simple yet powerful knowledge-distillation approach called interpretable mimic learning for interpretable deep models for clinical outcome predictions. They used an XGBoost model to learn the soft labels produced as outputs from the LSTM and deep neural networks. Our contribution through this project is to implement the same approach for the problem of early prediction of ICU readmissions by training models on Electronic Health Records (EHR) data.

III.Proposed Model

In this project we have implemented a *Knowledge Distillation* [11] approach, also called mimic learning [1], which involves training a large, slow, but accurate model and transfer its knowledge to a much smaller, faster, yet still accurate model. This architecture named interpretable mimic learning by [4] consists of a teacher model and a student model wherein teacher model learns complex features and transfers the knowledge to student model through their soft output labels. These soft labels are the real valued output of the teacher model, whose values ranges in [0, 1]. The intuition behind why this approach works is that soft labels from the teacher model usually contain more information than the original hard labels(0/1). These soft labels may also contain information pertaining to sequential relationships among features. When another model is trained using these soft labels, they may also learn the complex knowledge already learned by teacher model. The final predictions are made using student model's parameters learned from the knowledge of the teacher model. The teacher and student layers of this framework are discussed with more details in following section.

A. Teacher/Deep Layer

There are two types of features in our input data. The input data consists of time series data as well as static data. The LSTM model of the deep/teacher layer accepts the temporal EHR data X_T as input and produces soft labels y_{Ts} which are the predictors of whether the patient was readmitted to the ICU within 30 days. Considering the fact that LSTM performs well on sequential data, we have trained it using temporal features in addition to a DNN model which was trained on static features X_S to capture the inherent relationship between them and the state of being readmitted or not. The soft labels of LSTM y_{Ts} and DNN y_{Ss} are then combined using specific weights (based on accuracy) and then fed as output labels to the student model during training.

B. Student/Interpretable Layer

As the student model, we used XGBoost which is a gradient boosted tree. Instead of training it on hard labels, it is trained on soft labels y_S produced by the deep/teacher model as has been shown in 1.

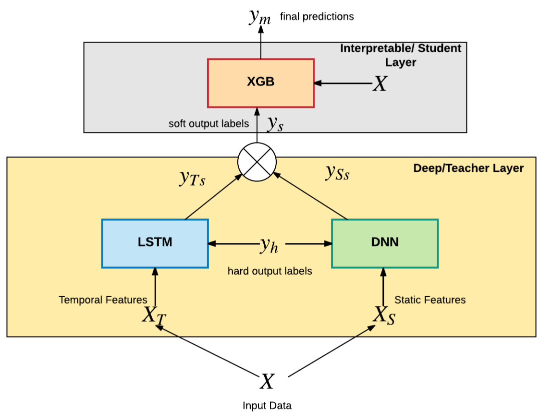

Fig. 1. Model Pipeline.

IV. DATA AND METHODS

A. Data Collection

The ICU data in MIMIC-III were collected between 2001 and 2012 at Beth Israel Deaconess Medical Center, Boston, MA, USA was used a data source for this analysis**??**. It consists of 58,000 de-identified hospital admissions for 38,645 adults and 7,875 neonates. There were 34005 unique patients with ICU admissions. Among the adults, 11.9% of the adults met our criterion of being readmitted within 30 days. Figure 2 shows the distribution of time differences between two consecutive admissions to ICU which resulted in readmission because of being less than 30 days.

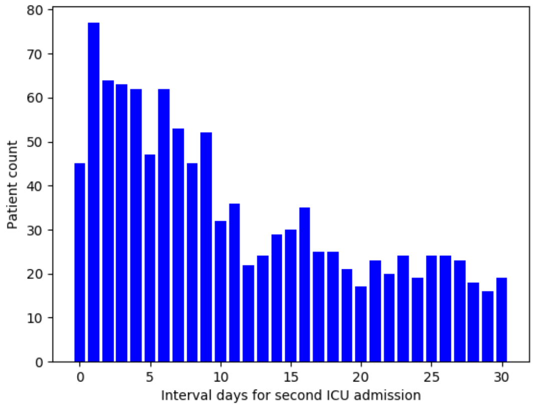

Fig. 2. Age distribution of patients with two or more ICU readmission.

We have analyzed the *Patients*, the *ICUStays*, the *Labevents* and *Chartevents* tables of MIMIC-III database for potential predictor variables.

1) Data Processing:

In preparing the dataset we came across many obstacles because of its huge size and our computational limitations of our systems. Hence, instead of working on entire feature range, we choose subset of features. For temporal features, we selected 17 variables generated from the *Chartevents* table of the MIMIC-III database as described by Harutyunyan,et al [10]. We generated time-series data files for each ICU stay ID (unique identifier for each ICU stay). We also identified outliers and corrected them within valid range specified by [10].

For static features, we selected 83 most frequently measured variables from the *Labevents* and the *Patients* tables of the MIMIC-III database. These features included demographic details e.g. age, gender marital status, lab results and severity scores e.g. SAPSII, SARS, OASIS. Static features from the *Labevents* table were generated by averaging values of temporal observations.

2) *Handling Missing Data:*

Transforming database tables into suitable time-sequenced files generated many missing values. These missing values were first imputed using forward filling and then mean-imputation afterwards. The intuition behind using forward-filling approach was that clinical variables are usually recorded at rates proportional to how quickly they are expected to change. So when a variable is absent, it is usually because clinicians believe it to be stable. [16].

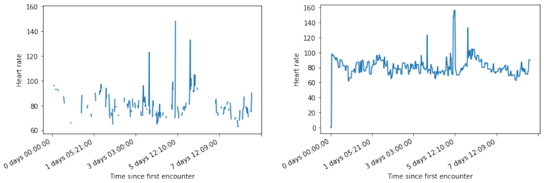

Fig. 3. Missing data imputation for Heart Rate (after normalization)

B. *Methods*

The pipeline consists of two major steps. First is training the teacher model and later training the student model based on the teacher model. Training teacher model itself consists of two parts. One of them is LSTM and the other one is DNN. These two models work on different parts of the data. LSTM works on the time series data while the DNN works on the static data such as age and gender. Once the Teacher model is trained, its soft labels are used as labels for training the student model. An advanced version of gradient boosting trees named XGBoost is used as the student model.

1) *LSTM:*

Long Short-Term Memory (LSTM) is a recurrent neural network for sequential data handling. It is used to avoid the vanishing gradient problem which is prevalent in other recurrent neural network architectures. Standard structure of an LSTM block contains input, forget and output gates. Temporal features are fed as input to this LSTM and soft label Yt is seen as the output.

2) *DNN:*

Deep neural network is simply a feedforward network with many hidden layers. Each layer contains multiple perceptron units with sigmoid activation functions. We have used two hidden layers for this project. Static features are fed as input to this DNN and soft label Ys is seen as the output.

The LSTM takes a sequence $\{x_t\}_{t\geq 1}^{T}$ of length T as its input and outputs a T-long sequence of $\{h_t\}_{t\geq 1}^{T}$ hidden state vectors using the following equations:

$$\begin{aligned}
i_t &= \sigma(x_t W_{xi} + h_{t-1} W_{hi} + w_{ci} \odot c_{t-1} + b_i) \\
f_t &= \sigma(x_t W_{xf} + h_{t-1} W_{hf} + w_{cf} \odot c_{t-1} + b_f) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(x_t W_{xc} + h_{t-1} W_{hc} + b_c) \\
o_t &= \sigma(x_t W_{xo} + h_{t-1} W_{ho} + w_{co} \odot c_t + b_o) \\
h_t &= o_t \odot \sigma_h(c_t)
\end{aligned}$$

3) *XGBoost:*

XGBoost uses Gradient boosting, which is a method which takes an ensemble of weak learners, usually decision trees, to optimize a differentiable loss function by stages. It combines weak learners into a single strong learner in an iterative fashion. XGBoost is short for Extreme Gradient Boosting. The two reasons which led us to use XGBoost were mainly its good execution speed as well as its better performance.

V. Experiment and Results

We performed binary classifications tasks on this dataset using the 3 models in the pipeline.

1) *LSTM Model:*

The model contained single LSTM layer with 256 neurons, dropout probability of 0.25 and sigmoid activations. We ran the model for 1 epoch with batch size of 8. Since, the dataset was huge, we split it into distinct files for each ICU stay ID and used Python Generator to fit the model on data batch-by-batch. Adam optimizer was used with learning rate of 0.001 and beta of 0.5. The LSTM model acheived an AUROC of 0.526 as shown in 4.

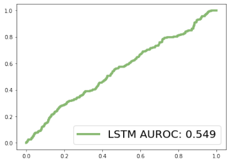

Fig. 4. AUROC curve for temporal features.

2) *DNN Model:*

The DNN model is composed of three layers with 80, 40 and 20 neurons respectively. We also used a dropout of 0.10, Adam as optimizer and sigmoid activation. The DNN model acheived AUROC of 0.705 as shown in 6.

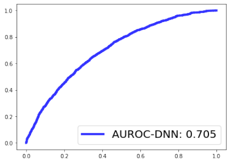

Fig. 5. AUROC curve for static features.

3) *XGBoost Model:*

We used grid searching to find the optimum hyper-parameters for the models. As our baseline model, XGBoost model was first trained on hard output labels (0,1) and achieved AUROC of 0.704. Then we trained XGBoost using soft labels to implement mimic-learning approach. XGBoost trained using soft labels achieved AUROC of 0.709. These soft labels were obtained by summing soft outputs from LSTM and DNN weighted by the accuracy of respective models.

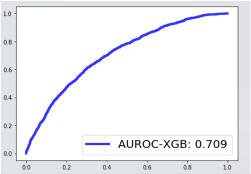

Fig. 6. AUROC curve for XGB trained from soft labels.

Model	**AUROC**
LSTM for temporal features	0.549
DNN for static feature	0.705
Baseline XGBoost	0.704
XGBoost with mimic learning	**0.709**

TABLE I. RESULTS FOR MODELS IN PIPELINE

4) *Interpretability:*

Our notion of interpretability is relative importance of predictor features. We used XGBoost to generate feature scores of different predictors fed into the model. XGBoost model generated following list of 10 most important predictors.

Important Features
RDW
Bicarbonate
PTT
Creatine Kinase (CK)
Base Excess
pO2
saps (severity score)
Urea Nitrogen
Eosinophils
Sodium, Whole Blood

TABLE II. LIST OF MOST IMPORTANT FEATURES FOR READMISSION PREDICTIONS

VI. CONCLUSION

We found that the performance of model trained on temporal features was not as good as that of the model which was trained on static features. One possible reason could be the irregularity in the time-sequences because the time-series data of different patients were not equally spaced so elements with specific indices of the time-series data could carry different information because of being captured at different times. Another reason could be the fact that we have included comparatively lesser number of predictor variables as temporal features. Analyzing the list of important features and feature scores, we realized that static features contributed more to the final prediction. This could also be because of lower accuracy of LSTM model compared to DNN model.

VII. FUTURE WORK

We believe that this study can further be improved by incorporating more predictors. We only included 17 important features from the chartevents table of the MIMIC-III database owing to computation limitations of our system. We also worked only on 83 features selected from labevents table. In future, we plan to include more features so that our deep model could learn more hidden relationships between predictors and readmission classification labels.

We also plan to introduce indicator variables to allow the LSTM to differentiate actual from missing or imputed measurements. We realized that high irregularities in the time-sequences of measurements. We hope binning time-sequences into regular sequences can further improve accuracy of LSTM models.

The MIMIC database also consists of notes on each patient. These notes contain valuable patient information which can be mined for prediction and further improve robustness of our prediction framework.

REFERENCES

[1] Jimmy Ba and Rich Caruana. “Do deep nets really need to be deep?” In: *Advances in neural information processing systems*. 2014, pp. 2654–2662.

[2] Sydney ES Brown, Sarah J Ratcliffe, and Scott D Halpern. “An empirical derivation of the optimal time interval for defining ICU readmissions”. In: *Medical care* 51.8 (2013), p. 706.

[3] Rich Caruana et al. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 1721–1730.

[4] Zhengping Che et al. “Interpretable deep models for icu outcome prediction”. In: *AMIA Annual Symposium Proceedings*. Vol. 2016. American Medical Informatics Association. 2016, p. 371.

[5] Edward Choi et al. “Doctor ai: Predicting clinical events via recurrent neural networks”. In: *Machine Learning for Healthcare Conference*. 2016, pp. 301–318.

[6] Edward Choi et al. “Retain: An interpretable predictive model for healthcare using reverse time attention mechanism”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 3504–3512.

[7] Carla A Chrusch et al. “High occupancy increases the risk of early death or readmission after transfer from intensive care”. In: *Critical care medicine* 37.10 (2009), pp. 2753–2758.

[8] Thomas Desautels et al. “Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach”. In: *BMJ open* 7.9 (2017), e017199.

[9] Cristóbal Esteban et al. “Predicting clinical events by combining static and dynamic information using recurrent neural networks”. In: *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*. IEEE. 2016, pp. 93–101.

[10] Hrayr Harutyunyan et al. “Multitask Learning and Benchmarking with Clinical Time Series Data”. In: *arXiv preprint arXiv:1703.07771* (2017).

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[12] Ulf Johansson et al. “Trade-off between accuracy and interpretability for predictive in silico modeling”. In: *Future medicinal chemistry* 3.6 (2011), pp. 647–663.

[13] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. “Visualizing and understanding recurrent networks”. In: *arXiv preprint arXiv:1506.02078* (2015).

[14] Phillip D Levin et al. “Intensive care outflow limitation-frequency, etiology, and impact”. In: *Journal of critical care* 18.4 (2003), pp. 206–211.

[15] Zachary C Lipton. “The mythos of model interpretability”. In: *arXiv preprint arXiv:1606.03490* (2016).

[16] Zachary C Lipton et al. “Learning to diagnose with LSTM recurrent neural networks”. In: *arXiv preprint arXiv:1511.03677* (2015).

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should i trust you?: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1135–1144.

[18] Andrew L Rosenberg and Charles Watts. “Patients readmitted to ICUs*: a systematic review of risk factors and outcomes”. In: *Chest Journal* 118.2 (2000), pp. 492–502.

[19] Norman Snow, Kathleen T Bergin, and Terrence P Horrigan. “Readmission of patients to the surgical intensive care unit: patient profiles and possibilities for prevention.” In: *Critical care medicine* 13.11 (1985), pp. 961–964.

[20] Evan G Wong et al. “Association of severity of illness and intensive care unit readmission: A systematic review”. In: *Heart & Lung: The Journal of Acute and Critical Care* 45.1 (2016), pp. 3–9.

[21] Jozef Zurada. “Could decision trees improve the classification accuracy and interpretability of loan granting decisions?” In: *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE. 2010, pp. 1–9.