2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Faster Healthcare Time Series Classification for Boosting Mortality Early Warning System

Yanke Hu[1] , Raj Subramanian[2], Wangpeng An[3], Na Zhao[4] and Weili Wu[5]

Abstract— **Electronic Health Record (EHR) and healthcare claim data provide rich clinical information for time series analysis. In this work, we provide a different angle of solving healthcare multivariate time series classification by turning it into a computer vision problem. We propose a Convolutional Feature Engineering (CFE) methodology, that can effectively extract long sequence dependency time series features. Combined with LightGBM, it can achieve the state-of-the-art results with 35X speed acceleration compared with LSTM based approaches on MIMIC-III In Hospital Mortality benchmark task. We deploy CFE based LightGBM into our Mortality Early Warning System at Humana, and train it on 1 million member samples. The offline metrics shows that this new approach generates better-quality predictions than previous LSTM based approach, and meanwhile greatly decrease the training and inference time.**

I. INTRODUCTION

Electronic Health Record (EHR) and healthcare claim data have skyrocketed over the past decade, due to prevalent adoption of internet and mobile technologies from hospitals and healthcare insurance companies. EHR contains rich text, visual and time series information such as a patient's medical and diagnose history, radiology images, etc which is the major source for managing a patient's health status. Healthcare claim data is the insurance claims that patients filed based on their health plans. EHR data normally contains more complete clinical information for patients, but one drawback is that different hospitals or medical systems may have different EHR formats, which leads to challenges of data integration. Compared to EHR, Healthcare claim data contains longitudinal information from all different parties in a patient-centered fashion. In the past, EHR and healthcare claim data are mainly used for patients' health status administration. Recently, there is an increasing interest for predictive analysis with EHR and healthcare claim data.

One important scenario in healthcare applications is multivariate time series classification. EHR contains rich short-term time-stamped nurse-verified physiological measurements for patients admitted to Intensive Care Unit (ICU), that can be utilized for in hospital mortality prediction, physiologic decompensation prediction, ICU length of stay prediction and so on. Healthcare claim data, on the other hand, can be used for longer term prediction such as 12-month mortality prediction for palliative care.

Traditional approaches for healthcare time series classification tasks heavily reply on feature engineering on timestamp attributes and then appending with task-specific classification or regression models [1]. Later, Recurrent Neural Network (RNN) approaches such as Long Short-Term Memory (LSTM) [2] was proven to be effective even when being trained with raw time series data without the need of feature engineering [3]. More recently, as RNN architectures being complained as less effective in parallel computing and slow, specific attention based modeling architectures were developed and evaluated to achieve the state-of-the-art result [4].

Despite more sophisticated model architectures emerging, the accuracy gain of these deep learning approaches is very slight. Moreover, the training and inference time of these sophisticated deep learning models is non-trivial. In this paper, we provide a different angle of solving healthcare multivariate time series classification by turning it into a computer vision problem. We propose a Convolutional Feature Engineering (CFE) methodology, that can effectively extract long sequence dependency time series features. Combined with LightGBM [5], a widely used Gradient Boosting Decision Tree (GBDT) method, it can achieve the state-of-the-art results with 35X speed acceleration compared with LSTM based approach on MIMIC-III In Hospital Mortality benchmark task [7]. We deploy CFE based LightGBM into our Mortality Early Warning System at Humana, and train it on 1 million member samples. The offline metrics shows that this new approach generates better-quality predictions than previous LSTM based approach, and meanwhile greatly decrease the training and inference time. The major contributions of this work are summarized as the following:

- We propose a different perspective of dealing with healthcare multivariate time series classification problem by encoding the vital signs into a $< 0, 1 >$ vector and aligning these vectors by time series. This will turn each time series sample into a 2-dimension image that can be applied with Convolutional Neural Networks (CNN) image classification models, which is much faster than RNN models.
- We propose a Convolutional Feature Engineering (CFE) methodology, that can effectively extract long sequence dependency time series features.Combined with LightGBM, we demonstrated that this approach can achieve

[1]Yanke Hu is a Senior Machine Learning Engineer with Humana, 2001 W John Carpenter Fwy, Irving, TX 75063, USA yhu@humana.com

[2]Raj Subramanian is a Lead Machine Learning Engineer with Humana, 2001 W John Carpenter Fwy, Irving, TX 75063, USA RSubramanian5@humana.com

[3]Wangpeng An is a Researcher with Tsinghua University, Beijing, China anwangpeng@gmail.com

[4]Na Zhao is a Researcher with Peking University School and Hospital of Stomatology, Beijing, China nanamozhao88@gmail.com

[5]Weili Wu is with Faculty of Computer Science Department ,University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, USA weiliwu@utdallas.edu

the state-of-the-art results with 35X speed acceleration compared with LSTM based approach on MIMIC-III In Hospital Mortality benchmark task.
- We deploy CFE based LightGBM into our Mortality Early Warning System at Humana, and train it on 1 million member samples. The offline metrics shows that this new approach generates better-quality predictions than previous LSTM based approach, and meanwhile decrease the training time from 33 hours to 1 hour.

II. Related Work

Traditional approach for mortality predictive analytics in ICU heavily involves the formulation of hand-crafted clinical decision rules (CDR) [8] , which suffers from questions of limitations of analytics insights, small preselected rules and constrained usability. Meanwhile, palliative care plays a more important role for old weak, ill or disabled, and automatic screening and notification will greatly help palliative team for proactively approaching the patients rather than relying on referrals from family physicians. The above two needs can be formulated as multivariate time series classification problems such as short-term ICU mortality prediction based on EHR data and longer term mortality prediction based on longitudinal healthcare claim data from data science perspective. Healthcare time series classification is very challenging because of irregular distribution of the sampling, wrong timestamp measurements and missing values. Recent year research shows that deep learning methods outperform traditional machine learning methods most of the time. Lipton et al. [3] demonstrated that a simple LSTM network with additional training strategies can outperform several strong baselines in 2015. With the growing need and interest of reproducing published methods to EHR, Medical Information Mart for Intensive Care (MIMIC-III) database [9] has established its reputation for evaluating different methods, because of its large size of de-identified clinical data of patients. From MIMIC-III dataset, Harutyunyan et al. proposed four clinical time series analysis tasks containing in hospital mortality prediction, physiological decompensation prediction, ICU length of stay prediction and 25-phenotype classification in 2017 [7]. They demonstrated the performance advantages of LSTM on these four tasks compared to traditional methods such as logistic regression, and joint training with LSTM on the four tasks will improve the performance further. In 2018, Song et al. proposed the sequence modeling architecture solely based on attention mechanism for multivariate time series classification. They demonstrated its performance improvement on the four MIMIC-III benchmark tasks [4].

III. Problem Formulation

A. Multivariate Time Series Classification

We denote a multivariate time series as F variables of length T

$X = (x_1, x_2, ..., x_T) \in \mathbb{R}^{T \times F}$

For each time stamp $t \in \{1, 2, ..., T\}$, $x_t \in \mathbb{R}^F$ represents the t-th measure of F variables, and x_t^f denotes the f-th variable of x_t. Both of the problems in this paper can be formulated as binary classification tasks, where we predict the label $l_i \in \{0, 1\}$ given the time series data $\mathbb{D}$, where $\mathbb{D} = \{(X_i)\}_{i=1}^N$, and $X_i = [x_1^{(i)}, ..., x_{T_i}^{(i)}]$

B. MIMIC-III In Hospital Mortality Prediction

This task is to predict the mortality risk from clinical time series variables recorded in the first 48 hours after the ICU admission. Here we use the MIMIC-III v1.4, which was released on September 2016. The database contains a cohort of 46520 unique patients from a total of 58976 admissions.We followed [7] to transform the data from original format into time series format. Each sample contains 48 time stamps of 17 vital signs including Capillary Refill Rate, Diastolic Blood Pressure, Fraction Inspired Oxygen, Glascow Coma Scale Eye Openning, Glascow Coma Scale Motor Response, Glascow Coma Scale Total, Glascow Coma Scale Verbal Response, Glucose, Heart Rate, Height, Mean Blood Pressure, Oxygen Saturation, Respiratory Rate, Systolic Blood Pressure, Temperature, Weight, pH. Our training dataset contains 17939 samples, validation dataset contains 3222 samples and test dataset contains 3236 samples. The ground truth label is determined by checking if the patient's date of death is between the ICU admission and discharge time. The overall mortality rate in the dataset is 11.60% (2830 of 24397 ICU stays). Since it's a very imbalanced labeled dataset, we use 3 metrics for the evaluation : (i) Area Under Receiver Operator Curve($AUROC$), (ii) Area Under Precision-Recall Curve($AUPRC$), and (iii) Minimum of Precision and Sensitivity($Min(Se, P+)$).

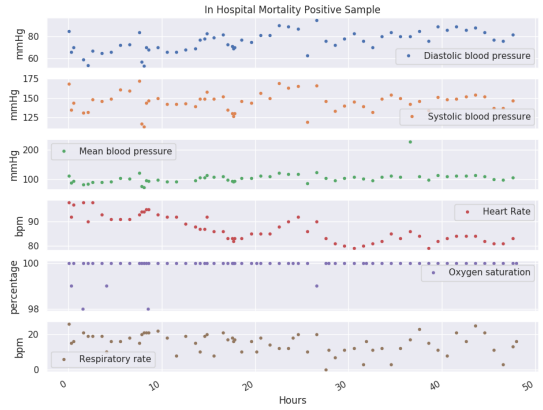
C. Humana Healthcare Claim One Year Mortality Prediction

This task is to predict the mortality risk of that patient within 12 months, given the Humana Claim data of that patient over the past 3 years. Here we use our Humana Mortality Early Warning benchmark dataset, which contains a cohort of 1 million unique patients. We standardized each patient's claim report into bi-monthly measure, so each sample contains $(3 \times 12 \times 2)$ time stamps of 100 selected vital symptom ICD9 codes including categories like Infectious and Parasitic Diseases, Neoplasms, Endocrine Diseases, Blood Organs, Mental Disorder, Nervous System, Circulatory System, Respiratory System, Digestive System, Genitourinary System, Skin and Subcutaneous Tissue, musculoskeletal System, etc. We have additional 10K patient samples as validation dataset, and additional 10k patient samples as test dataset. The overall mortality rate in the dataset is 17.70%, we use $AUROC$ for the evaluation.

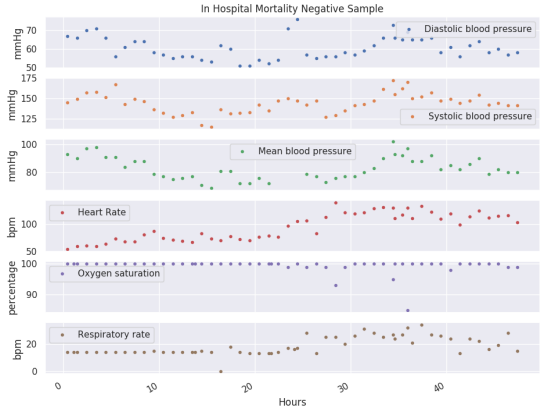
IV. Convolutional Feature Engineering

A. Existing Approaches

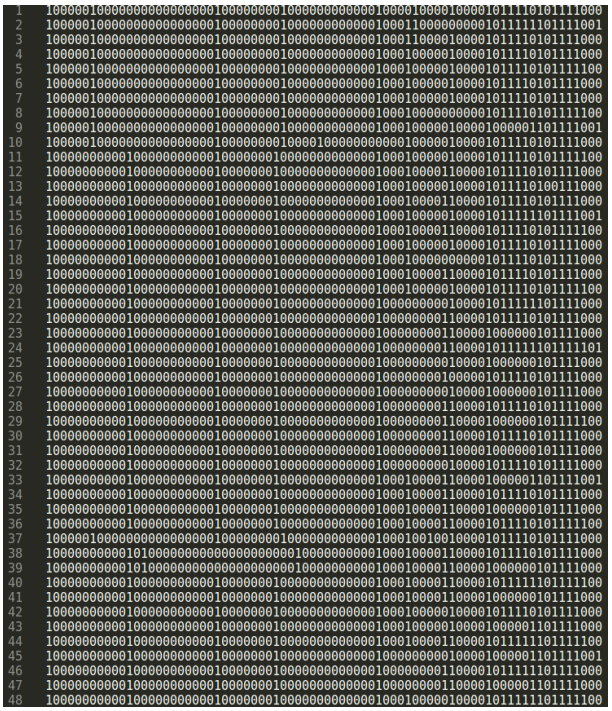
The past studies mainly utilize the Sub-Timeframe based feature engineering method with the logistic regression. In [7], for any given time series input sample, they compute six different sample statistic features (minimum, maximum, mean, standard deviation, skew and number of measurements) on seven different subsequences (full time series,



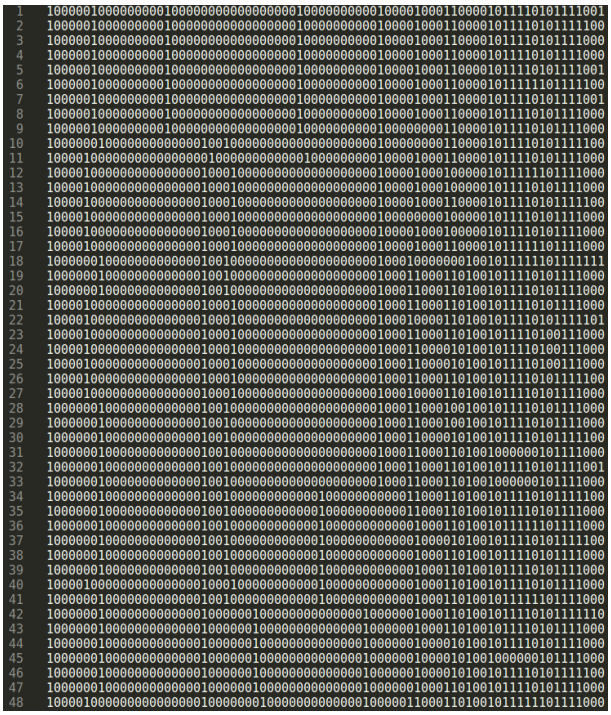
(a) Positive Sample by Vital Sign Measures and Time Series



(b) Negative Sample by Vital Sign Measures and Time Series



(c) Positive Sample after one-hot encoding



(d) Negative Sample after one-hot encoding

Fig. 1. MIMIC III In Hospital Mortality Data Sample

first 10% of time, first 25% of time, first 50% of time, last 50% of time, last 25% of time, and last 10% of time). Thus each time series input sample will generate $17 \times 7 \times 6$ features. One obvious problem of this method is that it mainly captures the statistic attribute of the data, but not the sequence dependency, so its accuracy performance is always outperformed by RNN approaches. Moreover, its running time is non-trivial. [7] then benchmarked a single layer 16 units LSTM model on the same 4 tasks and demonstrated that it can achieve a much better accuracy without the need of any feature engineering. RNN based approaches are good at capturing the long range dependencies of time series data, but they have their own problems: data has to be processed time stamp by time stamp both in training phase and inference phase, which negatively affect the speed performance.

B. Convolutional Neural Networks Approach

Here we are trying to look at this time series problem from a different angle. In the MIMIC-III In Hospital Mortality prediction case, we apply one-hot encoding of these 17 vital signs into a 76 length vector $\{x_i\}$ $(x_i \in \{0, 1\})$ (for non-categorical vital signs, we will first define degree levels and then transform them into degree categorical values). We align these vital sign vectors by the timestamp order, then each sample will turn into a two dimension $\{0, 1\}$ array. Now our original problem formulation will turn into:

Given the time series data $\mathbb{D}^{'}$, where $\mathbb{D}^{'} = \{(Y_i)\}_{i=1}^{N}$, $Y_i = [y_1^{(i)}, ..., y_T^{(i)}]$ and $y_t^f(i) \in \{0, 1\}$ denotes the f-th variable of

y_t, the task is to assign the label $l_i \in \{0, 1\}$ to Y_i.

Figure 1 (a) shows a positive MIMIC-III In Hospital Mortality data sample in vital sign time series format. Figure 1 (c) shows a positive MIMIC-III In Hospital Mortality data sample in $\{0, 1\}$ transformed format. Figure 1 (b) shows a negative MIMIC-III In Hospital Mortality data sample in vital sign time series format. Figure 1 (d) shows a negative MIMIC-III In Hospital Mortality data sample in $\{0, 1\}$ transformed format.

We then treat these two dimension arrays as images and apply CNN image classification models. We tested 3 light weight CNN models: Cifar_10 [13], SqueezeNet [11], MobileNetV2 [10]. The result shows that CNN based approaches can achieve the close to state-of-the-art result with 10 times faster speed than a single layer 16 units LSTM model, which we will elaborate more in the Experiment section.

C. Convolutional Feature Engineering

Since CNN approaches don't achieve the state-of-the-art result, here we apply with one more enhancement. We use CNN purely for feature engineering, and remove the last fully connected layer, which is responsible for classification task. We then feed these feature vectors into Gradient Boosting Decision Tree (GBDT) models, which are normally more suitable for structured data. Experiments showed that even a single convolution layer can reserve the sequence dependency attribute for not too large samples (e.g. MIMIC-III in hospital mortality prediction samples), and greatly speed up the GBDT training process, since it shrinks the feature vector length. Moreover, it's at least 200 times faster than traditional Sub-Timeframe feature engineering method on GPU servers.

V. Experiments

A. MIMIC-III in hospital Mortality Prediction

For MIMIC-III in hospital Mortality Prediction, our experiments were conducted on an on-premises server with hardware configuration: CPU Intel® Core™ i7-8700K CPU @ 3.70GHz × 12 , memory: 32 Gb , GPU: GeForce GTX 1080 Ti/PCIe/SSE2 .

Table 1 shows the speed comparison of different models on MIMIC-III In Hospital Mortality Prediction task. A single layer 16 units RNN model is generally 10 times slower than the three light CNN models. Gradient Boosting Decision Tree (GBDT) methods are even faster than the three light CNN models. Especially LightGBM is 2 times faster than XGBoost [12] on this task. The Sub-Timeframe feature engineering will generally speed up these two GBDT methods 3 times faster, while the 2×2 filter feature engineering will speed up these two GBDT methods 4 times faster, but Sub-Timeframe feature engineering is more than 200 times slower than 2×2 filter feature engineering.

Table 2 shows the best accuracy comparison of different models on MIMIC-III In Hospital Mortality Prediction task. These three light CNN models achieve close to RNN accuracy. An interesting observation is that even MobileNetV2 [10] is a more sophisticated network, it doesn't achieve better accuracy than classical Cifar_10 [13] and SqueezeNet [11]. 2×2 filter feature engineering not only speed up the GBDT methods, it also helps XGboost achieve the state-of-the-art $AUROC$, and helps LightGBM achieve the state-of-the-art $AUPRC$ and $min(Se, P+)$

Figure 2 shows the training process of different models on MIMIC-III In Hospital Mortality Prediction task. Figure 2 (a) shows training a single layer 16 units LSTM model for 100 epochs with dropout rate of 0.3 and batch size of 32. The model converges at the 25th epoch, so it takes around $25 \times 85 = 2125$ seconds to find the best model. Figure 2 (b) shows training classical Cifar_10 CNN network for 100 epochs with batch size of 32. The model converges at the 6th epoch, so it takes around $6 \times 2.87 = 17.22$ seconds to find the best model. Figure 3 (c) shows training XGboost with Sub-Timeframe feature engineering with learning rate of 0.07, number of estimators of 10000, max depth of 3 and early stopping rounds of 40. The model converges at the 230th epoch, so it takes around $230 \times 0.517 + 112 = 230.91$ seconds to find the best model. Figure 3 (d) shows training LightGBM with Sub-Timeframe feature engineering with learning rate of 0.07, number of estimators of 10000, number of leaves of 11, and the early stopping round of 80. The model converges at the 120th epoch, so it takes around $120 \times 0.172 + 112 = 132.64$ seconds to find the best model. Figure 3 (e) shows training XGboost with 2×2 filter feature engineering with learning rate of 0.07, number of estimators of 10000, max depth of 3 and early stopping rounds of 40. The model converges at the 250th epoch, so it takes around $250 \times 0.422 + 0.37 = 105.87$ seconds to find the best model. Figure 3 (f) shows training LightGBM with 2×2 filter feature engineering with learning rate of 0.07, number of estimators of 10000, number of leaves of 11, and the early stopping round of 80. The model converges at the 390th epoch, so it takes around $390 \times 0.156 + 0.37 = 61.21$ seconds to find the best model. Compared to the original LSTM model, LightGBM with 2×2 filter feature engineering can achieve the state-of-the-art result with around 35 times faster speed.

B. Humana Healthcare Claim One Year Mortality Prediction

For Humana Healthcare Claim One Year Mortality Prediction, our experiments were conducted on Microsoft Azure Standard_NC6s_v3 virtual machine. We trained 1 million samples with LightGBM with 2×2 filter feature engineering. The feature engineering time on this training dataset is around 72 seconds, and one training epoch takes around 11.2 seconds. The model converges at the 340th epoch, with $AUROC$ of 0.812 on the test dataset. so it takes around $340 \times 11.2 + 72 = 3880$ seconds ≈ 1.08 hours to find the best model. Our previous LSTM based model normally takes around 33 hours to find the best model, with $AUROC$ of 0.798 on the test dataset.

VI. Conclusions

Although the recent studies have shown that RNN based approaches are powerful in various time series use cases, yet its drawback of slow processing is easily neglected. In

Metrics	Training Epoch (s)	Inference on Test (s)	Feature Engineering (s)
LSTM(16 units, 1 layer)	85	4.292	0
GRU(16 units, 1 layer)	65	3.443	0
Cifar_10	2.87	0.355	0
SqueezeNet	3.98	0.782	0
MobileNetV2	8.65	1.002	0
XGBoost	1.899	0.212	0
LightGBM	0.781	0.3	0
XGBoost(Sub-Timeframe)	0.517	0.064	112
LightGBM(Sub-Timeframe)	0.172	0.09	112
XGBoost(2x2 filter)	0.422	0.048	**0.37**
LightGBM(2x2 filter)	**0.156**	**0.006**	**0.37**

TABLE I

SPEED COMPARISON FOR MIMIC-III IN HOSPITAL MORTALITY PREDICTION

Metrics	AUROC	AUPRC	min(Se, P+)
LSTM(16 units , 1 layer)	0.854	0.516	0.491
GRU(16 units , 1 layer)	0.851	0.500	0.482
Cifar_10	0.834	0.469	0.456
SqueezeNet	0.845	0.479	0.461
MobileNetV2	0.841	0.472	0.466
XGBoost	0.848	0.501	0.483
LightGBM	0.846	0.506	0.486
XGBoost (Sub-Timeframe)	0.847	0.478	0.471
LightGBM (Sub-Timeframe)	0.844	0.481	0.469
XGBoost (2x2 filter)	**0.856**	0.517	0.487
LightGBM (2x2 filter)	0.850	**0.523**	**0.508**

TABLE II

ACCURACY COMPARISON FOR MIMIC-III IN HOSPITAL MORTALITY PREDICTION

this work, we propose a different perspective of dealing with healthcare multivariate time series classification problem by encoding the vital signs into a $< 0, 1 >$ vector and thus turning it into a computer vision problem. We then propose a Convolutional Feature Engineering methodology, that can effectively reserve long sequence dependency time series features. Combined with LightGBM, it can achieve the state-of-the-art results with 35X speed acceleration compared with LSTM based approach on MIMIC-III In Hospital Mortality benchmark task. We deploy CFE based LightGBM into our Mortality Early Warning System at Humana, and train it on 1 million member samples. The offline metrics shows that this new approach generates better-quality predictions than previous LSTM based approach, and meanwhile greatly decrease the training and inference time. In the future, we will continue improving this approach and apply it to broader time series use cases for better palliative care.

ACKNOWLEDGMENT

We appreciate the encouraging comments from the reviewers. This work was supported by Humana and NSF of USA under grants 1747818 and 1907472.

REFERENCES

[1] Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ; ADHERE Scientific Advisory Committee, Study Group, and Investigators. 2005. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis.JAMA. 293:572–580. doi: 10.1001/jama.293.5.572.

[2] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

[3] Lipton, Z. C.; Kale, D. C.; Elkan, C.; and Wetzell, R. 2015.Learning to diagnose with lstm recurrent neural networks.arXiv preprint arXiv:1511.03677. .

[4] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias. 2018.Attend and Diagnose: Clinical Time Series Analysis using Attention Models, in AAAI

[5] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu, 2017. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 3149 - 3157

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP.

[7] Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; and Galstyan, A. 2017. Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:1703.07771.

[8] Taylor RA, Pare JR, Venkatesh AK, et al. 2016. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med;23:269–278.

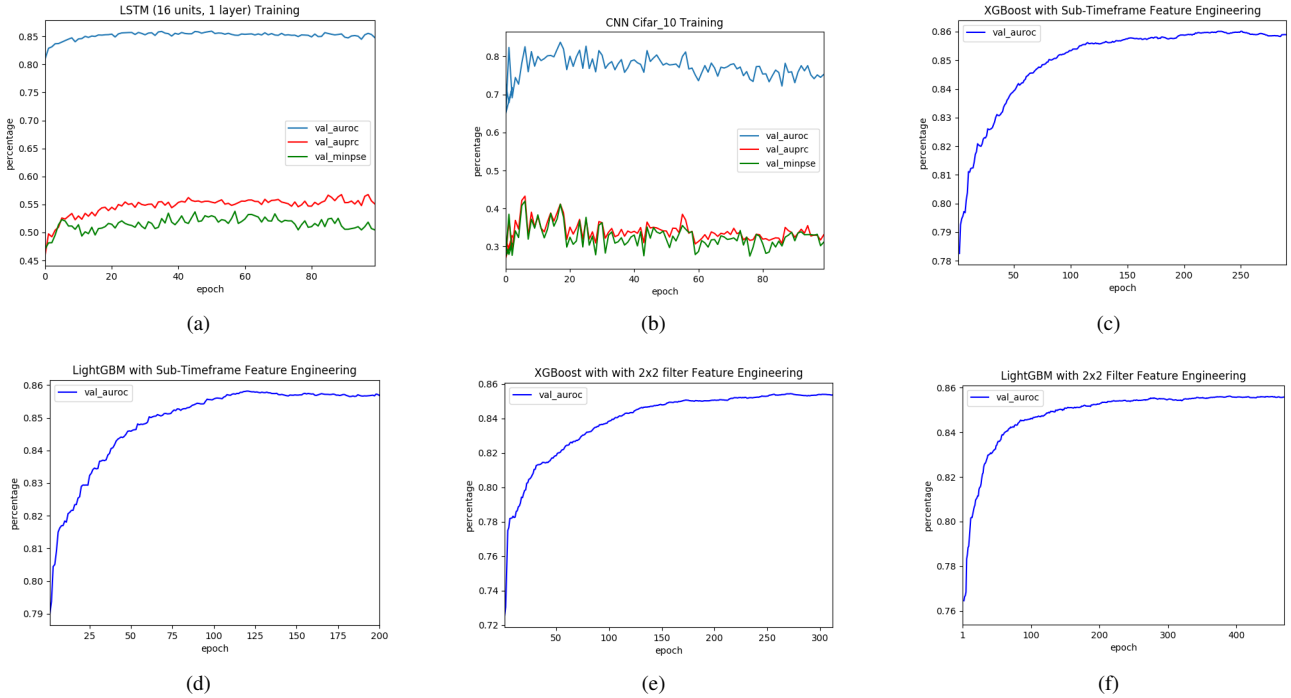


Fig. 2. MIMIC-III In Hospital Mortality Training

[9] Alistair E.W.Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi,Benjamin Moody, Peter Szolovits, Leo Anthony Celi,Roger G.Mark.2016.Mimic-iii, a freely accessible critical care database.Sci.data 3.

[10] Mark Sandler,Andrew Howard,Menglong Zhu,Andrey Zhmoginov,Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv preprint arXiv: 1801.04381,2018

[11] Forrest N. Iandola,Song Han, Matthew W. Moskewicz, Khalid Ashraf,William J. Dally,Kurt Keutzer.SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5MB model size.arXiv preprint arXiv: 1602.07360,2016

[12] T. Chen and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794. ACM

[13] $https://keras.io/examples/cifar10\_cnn/$