Title	Does Deep Learning REALLY Outperform Non-deep Machine Learning for Clinical Prediction on Physiological Time Series?
Author(s)	Liao, Ke; Wang, Wei; Elibol, Armagan; Meng, Lingzhong; Zhao, Xu; Chong, Nak Young
Citation	Proc. IEEE Int'l Conf. on Automation Science and Engineering, Lyon, France, August 23-27, 2021
Issue Date	2021-08
Type	Conference Paper
Text version	author
URL	http://hdl.handle.net/10119/18246
Rights	Copyright (C) 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Description	2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Lyon, France, August 23-27, 2021

Does Deep Learning REALLY Outperform Non-deep Machine Learning for Clinical Prediction on Physiological Time Series?

Ke Liao[1,2], Wei Wang[1], Armagan Elibol[2], Lingzhong Meng[3], Xu Zhao[4], and Nak Young Chong[2]

Abstract— Machine learning has been widely used in healthcare applications to approximate complex models, for clinical diagnosis, prognosis, and treatment. As deep learning has the outstanding ability to extract information from time series, its true capabilities on sparse, irregularly sampled, multivariate, and imbalanced physiological data are not yet fully explored. In this paper, we systematically examine the performance of machine learning models for the clinical prediction task based on the EHR, especially physiological time series. We choose Physionet 2019 challenge public dataset [1] to predict Sepsis outcomes in ICU units. Ten baseline machine learning models are compared, including 3 deep learning methods and 7 non-deep learning methods, commonly used in the clinical prediction domain. Nine evaluation metrics with specific clinical implications are used to assess the performance of models. Besides, we sub-sample training dataset sizes and use learning curve fit to investigate the impact of the training dataset size on the performance of the machine learning models. We also propose the general pre-processing method for the physiology time-series data and use Dice Loss [2] to deal with the dataset imbalanced problem. The results show that deep learning indeed outperforms non-deep learning, but with certain conditions: firstly, evaluating with some particular evaluation metrics (AUROC, AUPRC, Sensitivity, and FNR), but not others; secondly, the training dataset size is large enough (with an estimation of a magnitude of thousands).

I. INTRODUCTION

With the widespread adoption of Electronic Health Records (EHR), there is an increased emphasis on predictive models that can help clinical decisions and treatment recommendation. EHR contains various types of data, including demographics information, laboratory values, vital signs, physiology signals and treatment records, which are generally sparse, irregularly sampled, and multivariate. With the development of machine learning methods, especially deep learning and neural networks, more research focuses on clinical prediction or physiology management estimation with EHR using machine learning methods [3]. Deep learning shows excellent capabilities to extract information and features from big data, but the performance of deep learning

depends on the training dataset size, model structure, data quality, and similar other related conditions. Many research or actual work fails to reach the ideal target when utilizing deep learning models [4]. In some cases, non-deep learning, such as XGBoost has a similar or even better performance with deep neural networks [5].

In this paper, we systematically analyze the machine learning models' performance on the clinical prediction domain, using EHR, especially the physiology time series data. We propose a general framework to handle the raw multivariate time series data. We utilize the Physionet 2019 challenge public dataset [1] to predict Sepsis outcomes in ICU units, defined as a binary classification problem. And we train and evaluate the datasets with 10 baseline machine learning models: 3 deep learning models: (1) InceptionTime [6], (2) TCN (Temporal CNN) [7] and (3) Transformer [8] and 7 non-deep learning models: (4) LR (Logistic Regression), (5) SVM (Support Vector Machines), (6) KNN (K Nearest Neighbors), (7) GP (Gaussian Process), (8) DT (Decision Tree), (9) RF (Random Forests) and (10) XGBoost [5]. Then, we evaluate the performance of different machine learning models using 9 different evaluation metrics with specialized clinical meanings. Also, as the training dataset size is a crucial aspect for the models' performance, we sub-sample the training dataset and present a quantitative analysis of the relationship between the training dataset size and the performance of machine learning models using the learning curve fit method.

Based on the proposed systematic and quantitative analysis framework, we find that deep learning **indeed** outperforms non-deep learning only under certain conditions.

- The four evaluation metrics: AUROC, AUPRC, Sensitivity, and FNR show that deep learning outperforms non-deep learning. For example, AUROC is greater than 0.9 for deep learning models, while it is less than 0.8 for non-deep learning models.
- Labeled training dataset size: The larger training dataset improves the performance of the machine learning models, and deep learning models need a certain number of training dataset sizes to reach superior performance. The threshold value depends on different evaluation metrics. The overall estimate of training dataset sizes is about a magnitude of thousands.

II. RELATED WORK

There is an extensive body of clinical prediction based on EHR data, especially the physiology time series. Hatib *et al.* [9] use feature engineering to extract features from

[1]K. Liao and W. Wang are with the Ricoh Software Research Center (Beijing) Co., Ltd., Beijing, China (`{ke.liao, wei.wang5}@cn.ricoh.com`).

[2]K. Liao is a Ph.D. student of the Japan Advanced Institute of Science and Technology (JAIST). A. Elibol and N.Y. Chong are with the School of Information Science, JAIST, Ishikawa, Japan (`{s1820033, aelibol, nakyoung}@jaist.ac.jp`).

[3]L. Meng is with the Department of Anesthesiology, Yale University School of Medicine, New Haven, CT 06520, USA (`lingzhong.meng@yale.edu`).

[4]X. Zhao is with the Department of Anesthesiology, The Second Xiangya Hospital, Central South University, Changsha, Hunan Province, China (`zhaoxu109@hotmail.com`).

arterial pressure waveform to the prediction of an upcoming hypotensive event. Lee *et al.* [10] extract basic statistical features and then uses feedforward neural networks to predict mortality.

Some recent works try to extract the information or features from raw time series data, without feature engineering or statistical analysis. Kok *et al.* [7] use TCN for automated prediction of Sepsis outcome. Che *et al.* [2] develop LSTM for Sepsis prediction with missing values in raw data. Fawaz *et al.* [6] provide InceptionTime (a CNN-based model) to deal with time series classification problem. Zhao *et al.* [4] use the InceptionTime model to analyze the intraoperative time-series monitoring data, predicting the post-hysterectomy quality of recovery. Wu *et al.* [8] focus on the Transformer model and use it in the influenza prevalence prediction case.

III. METHODS

A. Data analysis and preprocessing

1) Data characteristics: This work is motivated by the analysis of physiological time series data in EHR, which are generally sparse, irregularly sampled, and multivariate. The EHR data contains the real-time physiological status of patients [3], such as vital signs or other physiology signals, laboratory values, demographics information, treatments, and outcomes. The data contains continuous and categorical data with varied time lengths and different sampling frequencies. The physiology time series are usually sparse, with missing data both in the time dimension and feature dimension. Also, the positive sample size is generally much smaller than the negative one in most clinical prediction domains. The imbalanced dataset is a fundamental problem in the clinical prediction field.

2) Public datasets: Physionet 2019 challenge dataset (early prediction of Sepsis from clinical data) includes hourly physiological time series data consisting of 8 vital signs, 26 laboratory values, and 6 demographic details of 40,336 samples from the MIMIC III database [1]. The time length varies from 8 to 336 time-steps. The outcome is the onset of sepsis.

3) Preprocessing:

a) Missing data: The physiology data is inconsistent in the timeline. Meanwhile, missing values are correlated with feature labels. As the physiology data reflect the clinical status of patients, we assume that the values vary little within a small time window. Thus, we use forward and backward fill for the missing points in the time domain. When the missing value relates to the feature labels, we fill other specific values with a special meaning, *e.g.*, -1 or N/A.

b) Various data length: Data length depends on time length and sampling frequencies. The time lengths of physiology time series are significantly different among features or patients. Also, the sampling frequencies are different among physiology features. The data filling process includes two steps: 1) firstly, utilize forward and backward filling according to the max sampling frequency. 2) fill each sample according to the max length of all samples with a special meaning, *e.g.*, -1 or N/A.

c) Continuous and categorical data: Generally, vital signs and laboratory values are continuous data and we use normalization to maintain the contribution of each feature and keep the model unbiased. Demographics information is mainly categorical data and we use one-hot encoding to convert the categorical data. After the filling, normalization, and one-hot encoding, processed EHR data is formatted for training and testing.

d) Imbalanced dataset: The imbalance issue is a common problem in the clinical domain: positive samples are much less than negative ones. For example, the Sepsis predction dataset has 40,336 samples with 2,932 positive patients, with a rate of 7.27%. Two main methods deal with the imbalance dataset: data augmentation and special loss functions. As the data augmentation methods affect the dataset size, we use Dice Loss [2] as the loss function of the prediction problem.

B. Models

1) Deep learning models: Deep learning models are based on neural networks that are composed of a large number of layers and parameters, which can be potentially used for extracting features from time series and downstream tasks such as regression, classification, forecasting, and feature embedding. We employ 3 different deep learning models to directly tackle the physiology time series for Sepsis prediction task, including InceptionTime [6], TCN (Temporal CNN) [7] and Transformer [8].

a) InceptionTime: InceptionTime is a 1D-CNN-based deep neural network. It consists of a series of Inception modules followed by a Global Average Pooling layer and a Dense layer with a softmax activation function. In each Inception module, multi-dimension time series data are transformed into 1D data. Three one-dimension filters with lengths of 10, 20, and 40, respectively, are applied simultaneously to the output. Residual blocks are added to mitigate the vanishing gradient problem.

b) TCN: TCN is a variant of the CNN and employs ordinary convolutions and dilations. It is suitable for sequential data with temporality and large receptive fields. The latent correlation among series can be learned from this model.

c) Transformer: The Transformer-based time series model consists of Transformer encoder layers. The network is composed of an input layer, a positional encoding layer, and a stack of transformer encoder layers. We use one transformer encoder layer in this work.

2) Non-deep learning models: For Non-deep learning models, we evaluate 7 models with the statistical information extracted from raw time series data. Some machine learning models are inconvenient to be used with raw time-series data, such as logistic regression. Usually, the preprocessing is needed to extract the basic (descriptive) statistical information or feature from the time-series data to be used in the machine learning methods [9],[10]. The statistical information includes the maximum, minimum, average, median, standard deviation, quantiles, moments, *etc.*

C. Training and Evaluation

1) Training process: We choose 36 features (8 vital signs, 26 laboratory values, and 2 demographic details) to predict the sepsis outcome as a binary classification. We divided the training dataset into training sets and test sets. The test sets have 4,033 samples (10% of all dataset sizes) and the training sets have 36,303 samples (90% of all dataset sizes). Also, the training sets are sub-sampled from 100 to 36,303 sample size: 100, 200, 400, 600, 800, 1000, 2000, 4000, 6000, 8000, 10000, 15000, 20000, 25000, 30000, 36303.

2) Performance metrics and evaluation: For the classification task, the most commonly reported metric is the area under the receiver operator characteristic curve (AUROC), which combines the information from sensitivity and specificity into a single value. We also report AUPRC since it is a good metric for describing model performance given imbalanced datasets. And we evaluate the accuracy, sensitivity (recall, TPR (true positive rate)), specificity (TNR, 1-FPR (false positive rate)), FNR, FPR, NPV, PPV, which represent actual clinical significance [11].

3) Training dataset size impact: The evaluation results of machine learning models are largely dependent on the training dataset size. We choose the learning curve approach [12] to explore the relationship between the performance and the training dataset size. The curve model is represented as an inverse power law function given by

$$y = f(x, \boldsymbol{b}) = 100 + b_1 \cdot x^{b_2}. \tag{1}$$

Using the evaluation results of the models trained with different training datasets sizes, we calculate the unknown parameters (b1 and b2) with nonlinear weighted least squares optimization using the following equation.

$$
\begin{aligned}
E(\boldsymbol{b}) &= \sum_{p=1}^{m} (\omega_p \cdot (t_p - y_p))^2 \\
&= \sum_{p=1}^{m} (\omega_p \cdot (t_p - f(x_p, \boldsymbol{b})))^2 \\
&= \sum_{p=1}^{m} (\omega_p \cdot r_p(\boldsymbol{b})^2) \\
&= \boldsymbol{R}^T W \boldsymbol{R},
\end{aligned}
\tag{2}
$$

where r_p is the residual between the real value and the fitted function.

IV. RESULT AND DISCUSSION

A. Comparison of machine learning models' performance

Fig. 1 contains the evaluation results of the Sepsis prediction task (a binary classification problem) based on different machine learning models.

Our findings include:

- The deep learning models outperform the non-deep learning methods for the AUROC, AUPRC, sensitivity, and FNR.

- For the accuracy, specificity, NPV, and FPR, both deep learning models and non-deep learning models have similar performance results.
- The performance of PPV evaluation metric differs from diverse machine learning models.
- All machine learning models could be divided into three groups based on performance: (1) deep learning models; (2) XGBoost, Random Forest, and Decision Tree; (3) Gaussian process, KNN, SVM, and logistic regression. The performance is ranked in the order of best to worst: Group 1, Group 2, Group 3.
- Among non-deep learning models, XGBoost performs best compared to the other non-deep learning models. XGBoost could not perform as well as deep learning methods though.

B. Training dataset size and learning curves

We also analyze the relationship between the training dataset size and the model performance. Fig. 2 depicts the impact of the training dataset size on the performance of the models, for the AUROC (upper left), AUPRC (upper right), Sensitivity (below left), and FNR (below right). Each line represents a model's performance according to the increasing number of training dataset sizes.

Our findings include:

- When the training dataset size is small, the deep learning models do not outperform the non-deep learning models, especially for the sensitivity evaluation metrics.
- With the increased training dataset size, the deep learning models show an excellent ability comparing with non-deep learning models for all four evaluation metrics.
- The increased number of training dataset size has a larger impact on Group 1 than Group 2, while no influence on Group 3.
- With the increased training dataset size, all the models' performance approach saturation.

V. CONCLUSIONS

Our systematical and quantitative study showed that deep learning indeed outperforms non-deep learning with specific training dataset size (an estimation of a magnitude of thousands) and some particular evaluation metrics (AUROC, AUPRC, Sensitivity, and FNR). No general effect of machine learning models or deep learning models was observed with small sample sizes (below estimation of a magnitude of thousands). Therefore, deep learning has the potential to be used as a tool for clinical prediction.

The future direction of our work will extend our quantitative analysis to investigate the relationship between the clinical research performance and the scale of the network, data points, the sampling frequency of raw data, augmentation of data, *etc.* It can assist researchers to determine the models and the dataset quality for clinical prediction research based on machine learning approaches.

Fig. 1. Performance comparison of machine learning models.
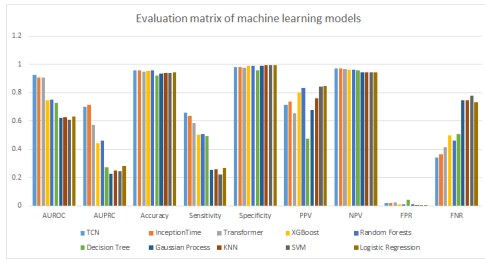
Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV	FPR	FNR
TCN	**0.924**	**0.698**	0.956	**0.659**	0.979	0.715	0.973	0.021	**0.341**
InceptionTime	**0.907**	**0.713**	0.957	**0.635**	0.982	0.735	0.972	0.018	**0.365**
Transformer	**0.909**	**0.572**	0.947	**0.584**	0.976	0.653	0.968	0.024	**0.416**
XGBoost	0.746	0.444	0.955	0.502	0.99	0.803	0.962	0.01	0.498
Random Forests	0.75	0.459	0.957	0.509	0.992	0.832	0.963	0.008	0.461
Decision Tree	0.726	0.271	0.923	0.495	0.957	0.474	0.96	0.043	0.505
Gaussian Process	0.622	0.226	0.937	0.253	0.991	0.679	0.944	0.009	0.747
KNN	0.625	0.248	0.94	0.256	0.994	0.758	0.945	0.006	0.744
SVM	0.609	0.244	0.94	0.222	0.997	0.844	0.942	0.003	0.778
Logistic Regression	0.631	0.279	0.943	0.266	0.996	0.848	0.945	0.004	0.734

(a) AUROC

(b) AUPRC

(c) Sensitivity

(d) FNR

Fig. 2. AUROC, AUPRC, Sensivitity and FNR Learning Curve of different models.

ACKNOWLEDGMENT

REFERENCES

[1] M. A. Reyna, C. Josef, S. Seyedi, R. Jeter, S. P. Shashikumar, M. B. Westover, A. Sharma, S. Nemati, and G. D. Clifford, "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[2] T. Vicar, P. Novotna, J. Hejc, M. Ronzhina, and R. Smisek, "Sepsis detection in sparse clinical data using long short-term memory network with dice loss," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[3] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[4] X. Zhao, K. Liao, W. Wang, J. Xu, and L. Meng, "Can a deep learning model based on intraoperative time-series monitoring data predict post-hysterectomy quality of recovery?" *Perioperative Medicine*, vol. 10, no. 1, p. 8, 2021. [Online]. Available: https://doi.org/10.1186/s13741-021-00178-4

[5] Y. Wang, B. Xiao, X. Bi, W. Li, J. Zhang, and X. Ma, "Prediction of Sepsis from Clinical Data Using LSTM and XGBoost," p. 4, 2019.

[6] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[7] C. Kok, V. Jahmunah, S. L. Oh, X. Zhou, R. Gururajan, X. Tao, K. H. Cheong, R. Gururajan, F. Molinari, and U. R. Acharya, "Automated prediction of sepsis using temporal convolutional network," *Computers in Biology and Medicine*, vol. 127, p. 103957, 2020.

[8] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," *arXiv preprint arXiv:2001.08317*, 2020.

[9] F. Hatib, Z. Jian, S. Buddi, C. Lee, J. Settels, K. Sibert, J. Rinehart, and M. Cannesson, "Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis," *Anesthesiology*, vol. 129, no. 4, pp. 663–674, 2018.

[10] C. K. Lee, I. Hofer, E. Gabel, P. Baldi, and M. Cannesson, "Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality," *Anesthesiology*, vol. 129, no. 4, pp. 649–662, 2018.

[11] H. B. Wong and G. H. Lim, "Measures of diagnostic accuracy: sensitivity, specificity, ppv and npv," *Proceedings of Singapore healthcare*, vol. 20, no. 4, pp. 316–318, 2011.

[12] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve

necessary high accuracy?" *arXiv preprint arXiv:1511.06348*, 2015.

REFERENCES

[1] M. A. Reyna, C. Josef, S. Seyedi, R. Jeter, S. P. Shashikumar, M. B. Westover, A. Sharma, S. Nemati, and G. D. Clifford, "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[2] T. Vicar, P. Novotna, J. Hejc, M. Ronzhina, and R. Smisek, "Sepsis detection in sparse clinical data using long short-term memory network with dice loss," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[3] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[4] X. Zhao, K. Liao, W. Wang, J. Xu, and L. Meng, "Can a deep learning model based on intraoperative time-series monitoring data predict post-hysterectomy quality of recovery?" *Perioperative Medicine*, vol. 10, no. 1, p. 8, 2021. [Online]. Available: https://doi.org/10.1186/s13741-021-00178-4

[5] Y. Wang, B. Xiao, X. Bi, W. Li, J. Zhang, and X. Ma, "Prediction of Sepsis from Clinical Data Using LSTM and XGBoost," p. 4, 2019.

[6] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[7] C. Kok, V. Jahmunah, S. L. Oh, X. Zhou, R. Gururajan, X. Tao, K. H. Cheong, R. Gururajan, F. Molinari, and U. R. Acharya, "Automated prediction of sepsis using temporal convolutional network," *Computers in Biology and Medicine*, vol. 127, p. 103957, 2020.

[8] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," *arXiv preprint arXiv:2001.08317*, 2020.

[9] F. Hatib, Z. Jian, S. Buddi, C. Lee, J. Settels, K. Sibert, J. Rinehart, and M. Cannesson, "Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis," *Anesthesiology*, vol. 129, no. 4, pp. 663–674, 2018.

[10] C. K. Lee, I. Hofer, E. Gabel, P. Baldi, and M. Cannesson, "Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality," *Anesthesiology*, vol. 129, no. 4, pp. 649–662, 2018.

[11] H. B. Wong and G. H. Lim, "Measures of diagnostic accuracy: sensitivity, specificity, ppv and npv," *Proceedings of Singapore healthcare*, vol. 20, no. 4, pp. 316–318, 2011.

[12] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" *arXiv preprint arXiv:1511.06348*, 2015.