ICU Mortality Prediction Using Long Short-Term Memory Networks

Manel Mili[1,3], Asma Kerkeni[2,3](✉), Asma Ben Abdallah[2,3], and Mohamed Hedi Bedoui[3]

[1] Faculty of Medicine, University of Monastir, Monastir, Tunisia
[2] Higher Institute of Computer Sciences and Mathematics, University of Monastir, Monastir, Tunisia
{manel.mili,asma.kerkeni,asma.benabdallah}@isimm.u-monastir.tn
[3] Laboratory of Technology and Medical Imaging, Faculty of Medicine, University of Monastir, Monastir, Tunisia
medhedi.bedoui@fmm.rnu.tn

Abstract. Extensive bedside monitoring in Intensive Care Units (ICUs) has resulted in complex temporal data regarding patient physiology, which presents an upscale context for clinical data analysis. In the other hand, identifying the time-series patterns within these data may provide a high aptitude to predict clinical events. Hence, we investigate, during this work, the implementation of an automatic data-driven system, which analyzes large amounts of multivariate temporal data derived from Electronic Health Records (EHRs), and extracts high-level information so as to predict in-hospital mortality and Length of Stay (LOS) early. Practically, we investigate the applicability of LSTM network by reducing the time-frame to 6-hour so as to enhance clinical tasks. The experimental results highlight the efficiency of LSTM model with rigorous multivariate time-series measurements for building real-world prediction engines.

Keywords: Electronic health record · Multivariate time-series data · MIMIC-III

1 Introduction

An ICU serves patients with severe complications or life-threatening injuries, which involve constant care in order to maintain normal bodily functions. To improve hospital services, it seems important to adequately select patients to be admitted to ICUs early on. In an ICU, the patient is monitored using Electronic Health Record (EHR) systems, entering many medical data a day including physiological measurements. Finding statistic models in these measurements has the potential to provide a high aptitude for more accurate and earlier predictions of future clinical events. This might not only help clinicians make more effective medical decisions but also facilitate an economical allocation of hospital

H. Yin et al. (Eds.): IDEAL 2022, LNCS 13756, pp. 242–251, 2022.
https://doi.org/10.1007/978-3-031-21753-1_24

resources. Naturally, mortality prediction and Length of Stay (LOS), are mainly performed with an interest in the prediction of possible outcomes, which are the death or survival of the patient, and for how long a patient may remain in the intensive units. Nevertheless, most available mortality and LOS prediction systems [1–4] in the literature were designed for at least 24-hour to provide a real-time or retrospective prediction on patients' mortality. To enhance prediction for early diagnosis, the main objective of this paper is to develop an end-to-end approach based on deep learning models, within a data mining framework, specifically intended for predicting mortality and LOS, based on multivariate time-series physiological measurements from the first few hours of admission, in particular after the first 6 h of a patient's acceptance in the ICU. The rest of the paper is organized as follows: Sect. 2 provides a comprehensive literature review on the state-of-the-art works. Section 3 details the process of dataset collection and preparation. Section 4 discuss the proposed model and presents its configuration and implementation tools. To consider the effectiveness of the proposed method, Sect. 5 deals with experiments. Ultimately, Sect. 6 concludes the paper and highlights the fundamental contributions.

2 Related Works

Over the past few decades, substantial researches are undertaken to affect predicting mortality risk and LOS tasks. A number of the more frequently used mortality prediction models in an ICU setting include SAPS-II [1] and SOFA [2]. SAPS-II was designed to estimate the probability of mortality, while SOFA was wont to describe organ dysfunction. Using the primary 24-hour patient physiological measurements, these scores are only designed to form one prediction. As a result, it's unknown how well each system predicts mortality following the primary day of admission. Moreover, it seems intuitively likely that straightforward clinical judgment also will discriminate more effectively as time passes. Existing tools are therefore slow to succeed in useful discriminatory effectiveness and aren't generally felt by clinicians to be useful to help decision-making once they will discriminate.

Adding to severity scores, several authors have converged on management mortality risk, as an example, Pirracchio et al. [4] aimed to develop a scoring procedure to predict mortality in ICUs supported Super-Learner (SL) model. They have proved that the SL method improved performance. However, the authors evaluated the performance of SL using data recorded within the primary 24-hour. Moreover, Darabi et al. [5] developed a model supported Gradient Boosted Tree (GBT) and Convolutional Neural Network (CNN) to estimate the mortality risk of patients admitted to ICUs. Their results prove usability a smaller number of features which will generate satisfactory outcomes for GBT, unlike, CNN that need a wealthy amount of knowledge for training. However, their model was designed within the period of 30-day after admittance.

In addition to mortality risk prediction, few researchers have converged to estimate LOS. Mentioning Gentimis et al. [6] who explored the utilization of

Neural Network (NN) for predicting the entire LOS of a patient within the hospital. The predictive model outperforms machine learning models. However, the studied scenarios considered time-frames > 5 days, or ≤ 5 days, to validate the potency of the model. Furthermore, Zebin et al. [7] applied an Auto-Encoder (AE) along side a dense neural network technique attempted at identifying short and long stays for patients. The proposed model improved the performance compared to employing a simplistic dense neural network for the classification task. However, their assessment results were validated using recordings observed after 24 h of admission.

To conclude, all the above-mentioned works only focused on predicting the risk of mortality and LOS for patients who required intensive care within a minimum of 24-hour of their ICU admission [3–5,8,9]. The challenge, therefore, lies within the early hours of a patient's admission, for instance, the primary 6 and 12 h. Additionally, not all critically ill patients can enjoy ICU admission. Hence, determining the priority of patients' treatments by the severity of their condition is crucial because the ICU is extremely costly with limited resources. The challenge, therefore, lies in triaging patients consistent with their medical conditions, while estimating their expected time of hospitalization. Adding to the present , most research has centered on the evaluation of the efficiency of their predictive models using univariate time-series data and that they didn't consider the potency of multivariate time-series records for improving the accuracy and therefore the efficiency of time-series modeling [10].

3 Dataset

This effort is conducted over the well-known publicly available, large-scale ICU database, the MIMIC-III [11], which presents a single-center electronic database developed by the MIT Lab for Computational Physiology, comprising health data related to 61.532 ICU admissions of 46.520 distinct de-identified patients admitted between 2012 and 2020.

3.1 Feature Engineering

Every day, different vital signs measurements are computed and analyzed during intensive stays. In this proceeding, we focused primarily, in hidden patterns within ICU time-series data and investigated the hypothesis that there is much useful knowledge in motifs within these data that can aid to improve prediction clinical tasks. This hypothesis is motivated by observations considered within several studies, for example, in [12], we found that in the event of a lack of oxygen transport, measurements in this time-frame of associated variables increase the risk of death. We therefore explored some temporal variables defined in acuity severity scoring systems and added others since they have proven to possess a powerful effect in predicting mortality and hence LOS [13]. These variables include "heart rate", "systolic BP", "diastolic BP", "mean BP", "respiratory rate", "oxygen saturation", "glasgow coma score", "blood urea nitrogen", "temperature", "white blood cells", and last not least "bilirubin".

Some of the foremost pertinent measurements could also be obtained using information available within the earliest phase [3]. So, we've extracted features for the primary 6 hours for every ICU stay. We have also extracted features for the 12 and 24 h so as to verify the effectiveness of the proposed model in maintaining its accuracy for long periods.

3.2 Feature Preprocessing

EHRs contain valuable information for estimating mortality risk and discharge time for ICU patients, but substantial missing and imbalanced data present mutual problems for the development and implementation of a prediction model. Hence, the subsequent two issues were identified and handled accordingly.

Missing Data Imputation. The percent of missing values for certain features is higher than 50%. To manage this problem, data imputation was performed including two strategies: we start by filling them using linear interpolation on each multivariate time-series data. Some observations are still missing after this imputation since there are missing data for certain variables. Hence, we impute missing observations using the Mean as the second strategy.

Imbalanced Data Regulation. The number of patients who passed away inside the intensive department is relatively small in comparison with the number of patients who survived, yielding an imbalanced dataset. To manage this problem, re-sampling methods were adopted since they are less sensitive to outliers than other techniques like Cost-sensitive classifiers [14] and Automatic support vector data description [15]. Two of the most common categories of re-sampling methods are under-sampling and over-sampling strategies. The former remove observations from the training dataset that belong to the dominant class, while the latter duplicate samples that belong to the lesser class, thus increasing its impact within the training process. We have applied the former on the dataset since the latter would make models inflexible in learning during the training process by causing overfitting. As a result, the size of the data was reduced from 33.6 Mo to 7.76 Mo, from 66, 7 Mo to 15 Mo and from 129 Mo to 29 Mo, over the 6-hour, 12-hour and 24-hour time-frames, respectively.

4 Methodology

The idea behind time-series prediction is to predict future events supported past values with reference to historical measurements and associated patterns. Turning to the philosophy of the research methodology, we would like to hold relevant information throughout the processing of medical data sequences, as physiological variables begin to decrease or increase over a period of your time, thus making it possible to predict future outcomes associated with patient conditions in care units. To reach these specific goals, a typical two-stage architecture is presented in Fig. 1.

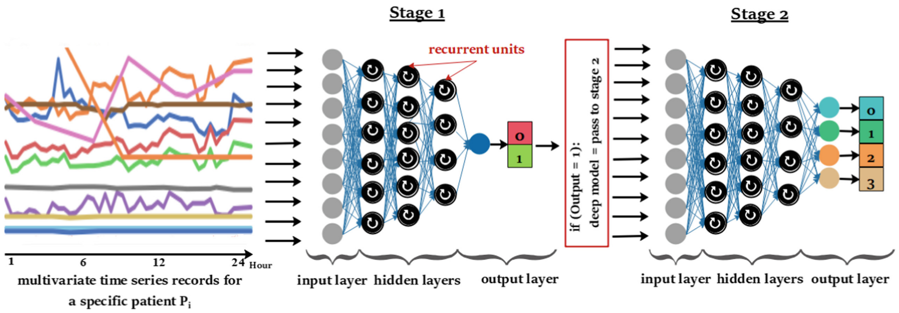


Fig. 1. A summary structure of two-stage architecture: within the first stage, a binary classifier is trained to predict mortality. Then, if the mortality is predicted to be positive, the model would further provide an estimation about LOS.

The philosophy behind the defined architecture is detailed as follows: we start by interpreting multivariate time-series of 11 past clinical records for each patient P_i:

$$P_i : X_{1,t_k}, X_{2,t_k}, ..., X_{11,t_k} \tag{1}$$

with $k = 1, \ldots, n$ and $n \in$ {6-hour, 12-hour, 24-hour}. In the first stage, a binary classifier is trained to predict the risk of mortality. In a mathematical interpretation we identify:

$$Class = \begin{cases} 0, & \text{survivors group} \\ 1, & \text{non-survivors group.} \end{cases}$$

Therefore, we define a knowledge set of two exclusion criteria: we start by filtering by $16 \leq$ age ≤ 89 [8]. Then, we exclude ICU stays of but one hour to get rid of obscurity in data due to unusual short stays. After filtering, we observe 49.632 ICU stays of 36.343 patients. While a multi-class classifier is trained at the second stage using a similar vital signs so as to predict LOS for those that are predicted dead in stage 1. Accordingly, we filter the ICU stays with death time ≤ 0. As a result, 5.718 in-hospital mortalities were obtained. We then label each data to at least one of the four classes represented below:

$$Class = \begin{cases} 0, & \text{if } death\_time\_hours < 6, \\ 1, & \text{if } 6 \leq death\_time\_hours < 12, \\ 2, & \text{if } 12 \leq death\_time\_hours < 24, \\ 3, & \text{otherwise.} \end{cases}$$

The proposed model will predict outcomes values by identifying short-term (6 h/12 h) and long-term (24 h) dependencies. For this purpose, we have employed the LSTM architecture [19]. This type of network improves the simple Multilayer Perceptron (MLP) network by including an output that depends on historical learned informations. The LSTM architecture is characterized by hidden units, called memory blocks. These units allow the network to remember information over short/long sequences. Moreover, these gates allow the LSTM model to beat the issues that inhibit the training of other deep models including RNNs and MLPs. This, and therefore the impressive results that may be achieved, are the rationale for its popularity on an outsized sort of problems [16,17].

4.1 Model Configuration

The efficient implementation of deep learning requires the selection and optimization of many hyperparameters, as well as extensive trial and error to find the optimal values. In order to assess the advanced performance, data is divided into training, test and validation sets; The training set is being used to train learning classifier; the validation set is used to fine-tune the parameters and estimate the behavior of the classifier; and the test set is going to be used to determine the efficiency of the classifier. Once data is splitted, we tune models using K-fold cross-validation. In this study, we set K = 3. The implemented LSTM model used Tanh activation function in the hidden layers and Sigmoid activation function in the output layer. Dropout with a rate of 0.2 is used as a regularization technique for weight optimization. In our model, a learning rate of $1e^{-03}$ is used, the number of epochs to train is set to 60 and the batch size is set to 100.

4.2 Model Implementation

In this work, the model was implemented using Keras framework, with TensorFlow backend. The implementation part of the proposed model consists of two stages:

1. Feature Engineering: we chose big data tool like Apache Hive 2.1.0 on Microsoft Azure remote cluster (2 head nodes and 1 worker node, each with 200 GB space, 14 GB RAM, and 4 processors), to perform data preprocessing and feature engineering.
2. Deep Learning using Colaboratory.

We also used Python and several packages for efficient model testing, hyperparameter tuning and model evaluation including: Pandas, NumPy, SciPy, Scikit-learn, Matplotlib, Seaborn.

5 Experimental Results

In this section, we describe the results of our experiments by evaluating the LSTM model against the traditional state of the art acuity scores and machine learning approaches that were used to predict possible future clinical events supported time-series measurements, including SOFA score, SAPS-II score, SL, SVM, LR, NB and CNN. Individual sets of parameters were tuned using 3-fold cross-validation to evaluate the potency of every fixed model. Experiments were conducted under three settings: using temporal physiological measures within 6-hour, 12-hour, and 24-hour time-frames. It's worth noting that SAPS-II and SOFA acuity scores use the primary 24 h of data to evaluate patient severity of illness.

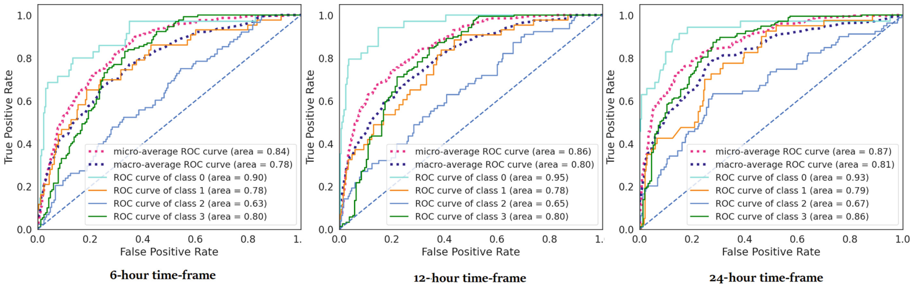
For binary-classifier, we opt for F1-score and MCC metrics to evaluate the effectiveness of the model. In gist, these two metrics were chosen because they provide a more realistic measure of a model's performance, and hence they are robust for binary classification problems [18].

Results outputs of different classifiers are presented in Table 1. In the light of the obtained results, fitting an LSTM model on the multivariate time-series records within a 6-hour time-frame has improved the prediction of early diagnosis of mortality risk for patients who remained in intensive departments. In fact, it is often seen from Table 1 that the LSTM model under the tuned configuration features a higher F1-score and MCC compare to the opposite mortality predictive approaches, which approved that the performance of the LSTM model is more consistent. Although the CNN model has attained a better F1-score and MCC within a 24-hour time-frame, the LSTM model outperformed it within 6-hour and 12-hour time-frames, validating its potency in predicting mortality risk as soon as possible following the admission of patients to the critical units.

Table 1. Mortality prediction performance for binary-classification approaches (The best performing model is highlighted in **bold**).

Classifier	Observation periods					
	6-hour		12-hour		24-hour	
	F1-score	MCC	F1-score	MCC	F1-score	MCC
SAPS-II	–	–	–	–	0.41	0.33
SOFA	–	–	–	–	0.06	0.14
NB	0.44	0.36	0.55	0.49	0.55	0.49
LR	0.09	0.20	0.14	0.25	0.17	0.28
SVM	0.33	0.20	0.30	0.18	0.27	0.17
SL	0.55	0.57	0.65	0.66	0.67	0.68
CNN	0.92	0.91	**0.97**	0.96	**0.99**	**0.99**
LSTM	**0.96**	**0.95**	**0.97**	**0.97**	0.96	0.96

Regarding multi-class classification, the average of the evaluation measures can provides a view on the overall results for the potency of LSTM fitted on the data aggregated over 6-hour within the prediction of LOS compared to those aggregated over 12-hour and 24-hour time-frames. Two major names to refer to averaged results are micro-average and macro-average. In gist, a macro-average will compute the metric independently for every class then take the average, whereas a micro-average will aggregate the contributions of whole classes to compute the average metric. Figure 2 summarizes Micro and Macro-average results for AUROC metrics and confirms that multivariate time-series data aggregated over a 6-hour time-frame offer rigorous multi-classification results compared with 12-hour and 24-hour time-frames that indicate slight improvement results.



6-hour time-frame

12-hour time-frame

24-hour time-frame

Fig. 2. ROC curves of the LSTM model fitted on data aggregated over 6-hour (in the left), 12-hour (in the middle), and 24-hour time-frames (in the right), applied for the multi-classification problem.

6 Conclusion and Future Works

Enhancing the excellence of care for patients and predicting future outcomes are the foremost important targets in critical care research. In this paper, and by deploying multivariate time-series data obtained from EHR-database MIMIC-III, we reveal that the LSTM model systematically outperforms all opposing predictive models of mortality using physiological measures observed during 6 and 12 h. These positive results recommend that access to the patient's physiological data trajectory as early as possible could enhance the potential in monitoring and predicting possible future events concerning the patient's conditions in ICUs. In future work, we arrange to apply the proposed model in other clinical tasks including early triage and risk assessment, prediction of physiologic decompensation, and identification of high-cost patients.

Acknowledgement. This work was supported by the Ministry of Higher Education and Scientific Research of Tunisia through the PEJC Young Researchers Encouragement Program (Project code 20PEJC 05-16).

References

1. Le Gall, J.R., Lemeshow, S., Saulnier, F.: A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA **270**(24), 2957–2963 (1993)
2. Simpson, S.Q.: New sepsis criteria: a change we should not make. Chest **149**(5), 1117–1118 (2016)
3. Awad, A., Bader-El-Den, M., McNicholas, J., Briggs, J., El-Sonbaty, Y.: Predicting hospital mortality for intensive care unit patients: time-series analysis. Health Inform. J. **26**(2), 1043–1059 (2020)
4. Pirracchio, R.: Mortality prediction in the ICU based on MIMIC-II results from the super ICU learner algorithm (SICULA) project. In: Secondary Analysis of Electronic Health Records, pp. 295–313. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43742-2_20
5. Darabi, H.R., Tsinis, D., Zecchini, K., Whitcomb, W.F., Liss, A.: Forecasting mortality risk for patients admitted to intensive care units using machine learning. Procedia Comput. Sci. **140**, 306–313 (2018)
6. Gentimis, T., Ala'J, A., Durante, A., Cook, K., Steele, R.: Predicting hospital length of stay using neural networks on mimic III data. In: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), pp. 1194–1201. IEEE (2017)
7. Zebin, T., Rezvy, S., Chaussalet, T. J.: A deep learning approach for length of stay prediction in clinical settings from medical records. In: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–5. IEEE (2019)
8. Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmark of deep learning models on large healthcare mimic datasets. arXiv preprint. arXiv:1710.08531 (2017)
9. Johnson, A.E., Dunkley, N., Mayaud, L., Tsanas, A., Kramer, A.A., Clifford, G.D.: Patient specific predictions in the intensive care unit using a Bayesian ensemble. In: 2012 Computing in Cardiology, pp. 249–252. IEEE (2012)
10. Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F.M., Preen, D.B., Stewart, L.M., Fatovich, D.M.: A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. J. Biomed. Inform. **57**, 62–73 (2015)
11. Johnson, A.E., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**(1), 1–9 (2016)
12. https://physionet.org/content/challenge-2012/1.0.0/
13. Vold, M.L., Aasebø, U., Wilsgaard, T., Melbye, H.: Low oxygen saturation and mortality in an adult cohort: the Tromsø study. BMC Pulm. Med. **15**(1), 9 (2015). https://doi.org/10.1186/s12890-015-0003-5
14. Perry, T., Bader-El-Den, M., Cooper, S.: Imbalanced classification using genetically optimized cost sensitive classifiers. In: 2015 IEEE Congress on Evolutionary Computation (CEC), pp. 680–687. IEEE (2015)
15. Sadeghi, R., Hamidzadeh, J.: Automatic support vector data description. Soft. Comput. **22**(1), 147–158 (2018). https://doi.org/10.1007/s00500-016-2317-5
16. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inform. Assoc. **24**(2), 361–370 (2017)

17. Reimers, N., Gurevych, I.: Optimal hyperparameters for deep lstm-networks for sequence labeling tasks (2017). arXiv preprint. arXiv:1707.06799
18. How to evaluate model performance in Azure Machine Learning Studio. https://docs.microsoft.com/fr-fr/azure/machine-learning/studio/evaluate-model-performance/
19. Lindemann, B., Müller, T., Vietz, H., Jazdi, N., Weyrich, M.: A survey on long short-term memory networks for time series prediction. In: Procedia CIRP, vol. 99, pp. 650–655 (2021)