Proceedings of the 21st International Conference on Global Research and Education (Inter-Academia 2024)

Biosignal-Based Machine Learning Predictors of Sepsis: A Mini-Review

M. Szumilas*

Faculty of Mechatronics, Institute of Metrology and Biomedical Engineering, Warsaw University of Technology, św. A. Boboli 8, 02-525 Warsaw, Poland

Doi: 10.12693/APhysPolA.146.388 *e-mail: mateusz.szumilas@pw.edu.pl

This work aims to provide insight into the most recent machine learning approaches to biosignal-based early sepsis prediction in the intensive care unit environment. A systematic search of the PubMed database revealed 29 original research papers. These works present sepsis prognosis and detection models that employ vital signs or densely sampled physiological waveforms (or their derivatives) acquired at the bedside or retrieved from electronic medical records. The papers were reviewed for the methods, predictors, datasets, number of participants, and performance achieved in the test set. Even though the sepsis prediction landscape is dominated by models that employ parameters derived from sparsely sampled biosignals, there are notable approaches built around densely sampled data, which speaks in favor of more synergistic solutions that benefit from both signal types. Given the already good quality of the models demonstrated using offline data, future research should prioritize achieving the promised performance in real-world intensive care unit operating conditions.

topics: early sepsis prediction, machine learning, intensive care unit (ICU), biosignals

1. Introduction

Sepsis is a life-threatening condition, and its early treatment improves patient outcomes. Artificial intelligence (AI) methods aimed at predicting incoming sepsis are tools with potentially large impact on critical care medicine in the near future [1, 2]. The models found in the literature are based on factors that may include patient demographics, lab test results, history of medical interventions, and biosignals (and their derivatives, like vitals). The signals themselves are commonly derived in already processed form from electronic medical records (EMRs), usually sampled once per hour. This review aims to delineate current trends in sepsis prediction in ICU settings within the subset of methods that employ machine learning, focusing on the approaches that take advantage of densely sampled biosignals (as relying on already processed data may reduce the amount of conveyed information). To support the planning of future studies in terms of prioritizing either further internal refinement of models or their intensive care unit (ICU) deployment, the predictors' readiness for introduction to real-world settings and the measures taken to guarantee their reliability were analyzed.

2. Materials and methods

PubMed base was searched for original research (published between 01.12.2020 and 30.11.2023) using the following query: "ICU AND (\"machine learning\" OR \"neural network\" OR \"reinforcement learning\" OR \"artificial intelligence\") NOT Review[PTYP] AND 2020/12/01:2023/11/30[PDAT]". In total, 1210 documents were identified and screened for works describing modeling approaches to prognosis or detection of sepsis in ICU patients. Only research papers were included. The methods had to be based on vital signs or physiological waveforms, or their derivatives, acquired at the bedside or retrieved from EMRs. Papers employing only laboratory or gene-related data were excluded. At this stage, 36 papers were preselected. Due to the low quality of the method description hindering research reproducibility, four papers were dropped from the analysis. Another three papers were dropped due to the lack of institutional access. Finally, 29 papers were included and reviewed for machine learning (ML) methods and variables used, dataset characteristics, diagnostic or prognostic type, and model maturity (development or external

validation). Where not given explicitly, the best model's F_1 score was calculated from precision and recall metrics.

3. Results

3.1. Databases

The majority of identified works used the following databases:

- Medical Information Mart for Intensive Care III and IV (MIMIC-III and MIMIC-IV): 10 and 2 papers, respectively [3–14];
- eICU Collaborative Research Database: 2 papers [5, 15];
- The PhysioNet/Computing in Cardiology Challenge 2019 database — PhysioNet/CinC 2019: 6 papers [8, 16–20];
- Other databases: 15 papers [5, 7, 13, 21–32].

Types of models proposed in the research papers (a single paper may belong to multiple classes):

- prognostic/diagnostic (i.e., sepsis prediction with 0 h horizon): 24/6;
- development/external validation phase: 29/4;
- infants as participants: 5;
- <18 years old patients as participants: 9.

3.2. Biosignals

The biosignals used for sepsis model development generally fall into the two following categories:

(i) physiological waveform-derived, densely sampled biosignals, as in 7 of the reviewed papers; modalities employed: electrocardiogram (ECG) [11, 13, 23, 24, 30, 32], pulse oximetry (POx) [23, 24], chest impedance (CI) [23, 30], and arterial blood pressure (ABP) [11, 13];

(ii) electronic health record (EHR)-derived, sparsely sampled biosignals (already processed and averaged), as used in 22 of the papers; modalities: blood pressure (systolic (SBP), diastolic (DBP), mean (MBP)), heart rate (HR), body temperature, oxygen saturation (SpO2), respiratory rate (RR), Glasgow Coma Scale subscores.

The densely sampled signals were processed to extract several parameters. For each of the works cited, the type of features derived were as follows:

- Stålhammar et al. [23]:
  - Inter-beat intervals (IBI) from ECG, respiratory rate (RR) derived by chest CI, peripheral SpO2 measured via POx, and body weight.
  - Summarized in 45-min windows using minimum, maximum, mean, standard deviation, skewness, and kurtosis.
  - For IBI, additionally, sample entropy (SampEn) and sample asymmetry were calculated.
- Kausch et al. [24]:
  - ECG-derived HR and POx-derived SpO2.
  - Summarized in 10-min windows using mean, standard deviation, skewness, kurtosis, max & min, and HR-SpO2 cross-correlation.
- Mollura et al. [11]:
  - Sixty-eight cardiovascular features from ECG and ABP.
  - Classical linear features extracted from normal-to-normal beats and spectral features computed from 5-min windows and successively averaged.
  - Nonlinear features calculated from the complete time series.
- Shashikumar et al. [13]:
  - Six features: standard deviation of RR intervals and MAP, average multiscale entropy of RR and MAP, and average multiscale conditional entropy of RR and MAP (calculated using 6-h sliding windows, with 5-h overlap).
- Liu et al. [29]:
  - Aggregation of raw data within the 1-min/5-min windows (upsampled when required).
  - Subsequent feature extraction from overlapping windows (15 min, 30 min, 1 h): mean, minimum, maximum, standard deviation, variance, skewness, and kurtosis (from all vital signs: HR, BP, RR, and SpO2).
  - Entropy was calculated from HR, BP, and RR in 1-h windows.
- Cabrera-Quiros et al. [30]:
  - Aggregation in 1-h windows.
  - Movement estimators, HR, and HRV features extracted from ECG.
  - Respiration features extracted from CI.
- Leon et al. [32]:
  - HRV features (time-domain, frequency-domain, nonlinear, and visibility graph indices) derived from 500 samples/s ECG recordings.

3.3. Performance in the test set

Most of the presented research papers evaluated proposed models on a hold-out test set. A few limited the evaluation to leave-one-out or multi-fold cross-validation. Results are shown in Table I.

TABLE I

Summary of the best results reported per research paper (when evaluated in the test set). The area under receiver operating characteristic (AUC) and F_1 score are provided in the context of dataset composition (N — number of patients in the dataset, SPr — sepsis prevalence) and prediction horizon in hours (Tsep). Models that employ densely sampled biosignals are marked with *.

Best model	N	SPr	AUC	F_1	Tsep [h]
XGBoost	2 000	50%	0.91	0.32	6
DNN, GRU	27 189	23.8%	0.788	0.285	24
XGBoost	47 185	5.8%	0.745	n/a	≥ 24
Deep model + self-attention	136 478	18.8%	0.761	0.447	1.7
RL + LSTM	40 336	7.3%	0.911	0.467	24
Naïve Bayes classifier*	378	9.5%	0.67	n/a	24
XGBoost	4 603	26%	0.88	n/a	4
			0.83	n/a	24
RF	40 336	7.3%	n/a	0.87	12
XGBoost*	2 494	11%	0.799	n/a	24
GBM + TrL	7 344	16%	0.93	n/a	4
MLP + TrL	7 344	16%	0.93	n/a	4
XGBoost	27 040	16.3%	0.825	0.165	0
XGBoost	2 932	n/a	0.862	0.111	0
RNN + GA	31 575	4.8%	0.94	n/a	3
VAE(GMM)	18 814	38%	0.82	0.652	3
LSTM			0.80	0.660	3
biLSTM	24 219	3.8%	0.768	n/a	2
			0.739	n/a	4
			0.761	n/a	6
SVM with linear basis*	142	50%	0.92	0.83	0
Logit*			0.91	0.85	0
CNN	2 893	20%	0.84	0.597	3
			0.85	0.531	0
DNN with encoder	515 720	7.1%	0.953	0.537	12.2
MLP	113	67%	n/a	1	0
RFs	3 031	47.2%	0.824	n/a	0
Ensembles of MA-ARMA + DFs	3 298	13.5%	0.975	0.864	0
LightGBM + PMI factorization	40 336	7.3%	0.862	0.164	6
GRU + NN*	25 820	5.6%	0.90	n/a	4
RF, NN + statistical analysis*	882	50%	n/a	0.74	6
	634	50%	n/a	0.71	10
Logit*	64	50%	0.79	0.799	3
NLP + RF + Logit + voting ensemble	5 317	6.15%	0.94	0.869	0
			0.90	0.805	24
			0.94	0.87	12
			0.92	0.849	6
			0.92	0.834	4
LSTM/GRU + CNN	40 336	6%	0.92	0.717	4
			0.87	0.652	8
			0.84	0.629	12
RFs + CNN	5 154	27.2%	0.972	n/a	6
			0.982	n/a	0
GA + RUSBoost	40 336	7.3%	0.843	n/a	6
Logit*	49	49%	0.877	n/a	6

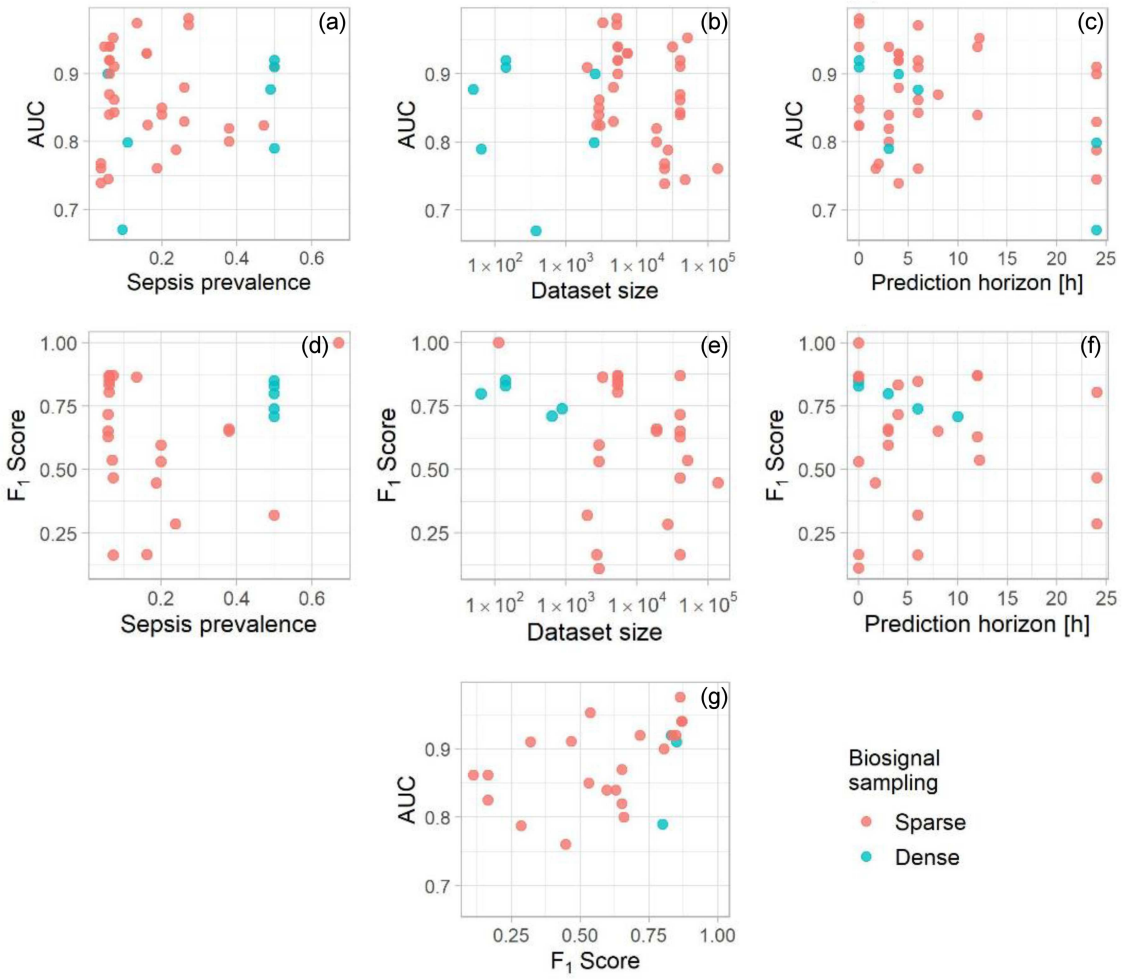

Fig. 1. Model performance metrics (AUC, F_1 score) presented in relation to the dataset parameters (number of patients and sepsis prevalence) and the prediction horizon.

As the performance varied with a prediction horizon between 0 h (diagnosis) and 24 h (prognosis), more than one result from a single paper is reported when possible. The following model-related abbreviations are used:

NN — neural network,
CNN — convolutional NN,
DNN — deep NN,
RNN — recurrent NN,
DF — deep forest,
GA — genetic algorithm,
GBM — gradient boosting machine,
GRU — gated recurrent unit,
Logit — logistic regression model,
LSTM — long short-term memory,
MA-ARMA — multi-activations autoregressive moving average,
MLP — multilayer perceptron,
NLP — natural language processing,
PMI — pointwise mutual information,
RF — random forest,
RL — reinforcement learning,
SVM — support vector machine,
TrL — transfer learning,
VAE(GMM) — variational autoencoder (Gaussian mixture model),
XGBoost — eXtreme gradient boosting.

4. Discussion

The majority of models with the best scores reported per paper are, to some extent, employing neural networks (15 out of 32 models) or tree-based methods (13 out of 32), whereas the final model commonly merges different ML approaches (Table I). No evident relationship was observed between performance and dataset characteristics (size and sepsis prevalence) or prediction horizon (Fig. 1a–f). Some models, having relatively high AUC reported, presented low F_1 scores (Fig. 1g). This may occur in imbalanced datasets where the AUC cannot show actual performance, as revealed

by F_1 [33]. Underreporting the latter raises concerns about the tested model's error rate in the minority class (most often sepsis cases). A precision–recall curve is an example of a characteristic worth presenting in such cases.

4.1. Densely vs sparsely sampled biosignals

Employing densely sampled biosignals should improve the dynamical properties of a prediction model, contrary to the dampening effect evoked by the inclusion of constant or slowly changing factors, e.g., demographic variables [24]. In particular, dense sampling is arguably the only way to acquire enough data to get a model prediction as early as possible for patients arriving at the healthcare facility, which is the use case for rapid sepsis detection models [11], commonly sought after in neonatal settings.

4.2. Transferability issues

The lack of models' generalization capabilities emerges when these models face data shift due to differences in the monitored population, specifically covariate shift (change in predictors distributions), prior probability shift (change in outcomes), or concept shift (change in the underlying relationship) [15]. Models that deteriorate on external data can be fine-tuned with such techniques as transfer learning to improve their performance in new datasets [7]. Continuous, rather than one-time, adaptation may be required as a part of long-term model maintenance.

4.3. Class imbalance

With the estimated 6% sepsis prevalence in hospitals [31], researchers are faced with large datasets being dominated by non-sepsis patients. This needs addressing during model training. One approach is to employ adequate algorithms, like the RUSBoost, following gradient-boosting schemes [20]. Another is to use sampling techniques, like the synthetic minority oversampling technique (SMOTE) [31], or a mixture of oversampling of sepsis-related data and undersampling of controls [9]. Finally, a balanced dataset may be synthesized based on clinical knowledge [3].

4.4. Real-world evaluation

As AI is introduced in critical care, clinicians encounter many published models that are not accompanied by real-world validation [34]. Indeed, the post-development model validation on real-world data was conducted in only one of the reviewed papers [7]. Moreover, the performance measures typically calculated during model testing might not be enough for the decision-makers responsible for introducing early sepsis prediction models to clinical practice. For example, more adequate analyses may comprise in situ observation of the long-term benefits for patients [35] or even investigation of the potential cost and cost-effectiveness impact [36] due to the implementation of ML-based sepsis prognostic systems.

5. Conclusions

The ICU ML and biosignal-based sepsis prediction landscape is dominated by models that employ parameters derived from sparsely sampled signals. This reduction of data complexity happens at the expense of potentially losing much of the information conveyed by signal dynamics in the small timescale. There are, however, approaches built around densely sampled data, commonly in the neonatal setting, aimed at sepsis prediction, even up to 24 hours before onset. To achieve fast-responding models that simultaneously keep track of the patient's medical history, the path to follow is arguably in between where combining static factors, slowly-changing vitals, and densely sampled biosignals yields a synergistic effect. However, even without further adjustments of the input variables, it should be admitted that most reviewed models show good prediction quality at the current development stage. Considering the importance of advancing the discussed solutions towards the phase of actual implementation, future research should now focus not on the model refinement using the offline data, but rather on reaching the promised performance in the real-world operating conditions of ICUs. This may include investigating the feasibility of continuous model tuning or studying methods that prevent the deterioration of the prediction when faced with missing or corrupt input data.

Acknowledgments

The author would like to thank Gerard Cybulski for inspiring discussions on the topic.

References

[1] Z. Yang, X. Cui, Z. Song, *BMC Infect. Dis.* **23**, 635 (2023).

[2] M.R. Pinsky, A. Bedoya, A. Bihorac et al., *Crit. Care* **28**, 113 (2024).

[3] S. Lyra, J. Jin, S. Leonhardt, M. Lüken, in: *2023 45th Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2023.

[4] Z. Jiang, L. Bo, L. Wang, Y. Xie, J. Cao, Y. Yao, W. Lu, X. Deng, T. Yang, J. Bian, *Comput. Methods Programs Biomed.* **241**, 107772 (2023).

[5] M. Moor, N. Bennett, D. Plečko, M. Horn, B. Rieck, N. Meinshausen, P. Bühlmann, K. Borgwardt, *EClinicalMedicine* **62**, 102124 (2023).

[6] J. Li, F. Xi, W. Yu, C. Sun, X. Wang, *JMIR Form. Res.* **7**, e42452 (2023).

[7] Q. Chen, R. Li, C. Lin et al., *BMC Med. Inform. Decis. Mak.* **22**, 343 (2022).

[8] S. Liu, W. Wang, M. Liu, X. Sun, *IEEE J. Biomed. Health Inform.* **26**, 4258 (2022).

[9] J.K. Kim, W. Ahn, S. Park, S.-H. Lee, L. Kim, *Int. J. Environ. Res. Public Health* **19**, 2349 (2022).

[10] G. Ramos, E. Gjini, L. Coelho, M. Silveira, in: *2021 43rd Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, 1916.

[11] M. Mollura, L.-W.H. Lehman, R.G. Mark, R. Barbieri, *Philos. Trans. A Math. Phys. Eng. Sci.* **379**, 20200252 (2021).

[12] I. Persson, A. Östling, M. Arlbrandt, J. Söderberg, D. Becedas, *JMIR Form. Res.* **5**, e28000 (2021).

[13] S.P. Shashikumar, C.S. Josef, A. Sharma, S. Nemati, *Artif. Intell. Med.* **113**, 102036 (2021).

[14] T. Aşuroğlu, H. Oğul, *Comput. Methods Programs Biomed.* **198**, 105816 (2021).

[15] J. Gao, P.L. Mar, G. Chen, *AMIA Jt. Summits Transl. Sci. Proc.* **2021**, 220 (2021).

[16] H. Dai, H.-G. Hwang, V.S. Tseng, *IEEE J. Biomed. Health Inform.* **27**, 3610 (2023).

[17] E.A.T. Strickler, J. Thomas, J.P. Thomas, B. Benjamin, R. Shamsuddin, *Sci. Rep.* **13**, 3067 (2023).

[18] N. Nesaragi, S. Patidar, V. Aggarwal, *Comput. Biol. Med.* **134**, 104430 (2021).

[19] A. Rafiei, A. Rezaee, F. Hajati, S. Gheisari, M. Golzan, *Comput. Biol. Med.* **128**, 104110 (2021).

[20] N. Nesaragi, S. Patidar, *Crit. Care Med.* **48**, e1343 (2020).

[21] R. Liu, J. Greenstein, J.C. Fackler, J. Bergmann, M.M. Bembea, R.L. Winslow, *Crit. Care Explor.* **3**, e0442 (2021).

[22] T. Kim, Y. Tae, H.J. Yeo et al., *J. Clin. Med.* **12**, 7156 (2023).

[23] A.M. StÍlhammar, A. Honoré, K. Adolphson, D. Forsberg, E. Herlenius, K. Jost, *Acta Paediatr.* **112**, 1443 (2023).

[24] S.L. Kausch, J.G. Brandberg, J. Qiu et al., *Pediatr. Res.* **93**, 1913 (2023).

[25] M. Sung, S. Hahn, C.H. Han, J.M. Lee, J. Lee, J. Yoo, J. Heo, Y.S. Kim, K.S. Chung, *JMIR Med. Inform.* **9**, e26426 (2021).

[26] S.P. Shashikumar, G. Wardi, A. Malhotra, S. Nemati, *NPJ Digit. Med.* **4**, 134 (2021).

[27] L.A. Arriaga-Pizano, M.A. Gonzalez-Olvera, E.A. Ferat-Osorio et al., *Comput. Methods Programs Biomed.* **210**, 106366 (2021).

[28] X. Chen, R. Zhang, X.-Y. Tang, *Eur. Rev. Med. Pharmacol. Sci.* **25**, 4693 (2021).

[29] Z. Liu, A. Khojandi, A. Mohammed, X. Li, L.K. Chinthala, R.L. Davis, R. Kamaleswaran, *Comput. Biol. Med.* **131**, 104255 (2021).

[30] L. Cabrera-Quiros, D. Kommers, M.K. Wolvers et al., *Crit. Care Explor.* **3**, e0302 (2021).

[31] K.H. Goh, L. Wang, A.Y.K. Yeow, H. Poh, K. Li, J.J.L. Yeow, G.Y.H. Tan, *Nat. Commun.* **12**, 711 (2021).

[32] C. Leon, G. Carrault, P. Pladys, A. Beuchée, *IEEE J. Biomed. Health Inform.* **25**, 1006 (2021).

[33] L.A. Jeni, J.F. Cohn, F. De La Torre, in: *2013 Humaine Association Conf. on Affective Computing and Intelligent Interaction*, IEEE, 2013, p. 245.

[34] V. Bellini, M. Valente, P. Pelosi, P. Del Rio, E. Bignami, *Neurocrit. Care* **37**, 170 (2022).

[35] M. Schootman, C. Wiskow, T. Loux, L. Meyer, S. Powell, A. Gandhi, A. Lacasse, *J. Crit. Care* **71**, 154061 (2022).

[36] O. Ericson, J. Hjelmgren, F. Sjövall, J. Söderberg, I. Persson, *J. Health Econ. Outcomes Res.* **9**, 101 (2022).