2022 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-6654-8485-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISCAS48785.2022.9937902

Real-time sepsis prediction using fusion of on-chip analog classifier and electronic medical record

Sudarsan Sadasivuni*, Monjoy Saha†, Sumukh Prashant Bhanushali§, Imon Banerjee‡, and Arindam Sanyal§
*Department of Electrical Engineering, University at Buffalo, Buffalo, NY 14260, USA.
†Department of Bio-Medical Informatics, Emory University, Atlanta, GA.
‡Mayo Clinic, Phoenix, AZ 85054, USA.
§School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA.
Email: ssadasiv@buffalo.edu

***Abstract*—This work presents a fusion artificial intelligence (AI) framework that combines patient electronic medical record (EMR) and physiological sensor data to accurately predict early risk of sepsis 4 hours before onset. The fusion AI model has two components - an on-chip AI model that continuously analyzes patient electrocardiogram (ECG) data and a cloud AI model that combines EMR and prediction scores from on-chip AI model to predict fusion sepsis onset score. The on-chip AI model is designed using analog circuits for high energy efficiency that allows integration with resource constrained wearable device. The on-chip AI reduces by $4.5\times$ compared to digital baseline, and by $4\times$ compared to state-of-the-art bio-medical AI ICs. Combination of EMR and sensor physiological data improves prediction performance compared to EMR or physiological data alone, and the late fusion model has an accuracy of 92.2% in predicting sepsis 4 hours before onset. The key differentiation of this work over existing sepsis prediction literature is the use of single modality patient vital (ECG) and simple demographic information, instead of comprehensive laboratory test results and multiple vital signs.**

keywords- Sepsis prediction, on-chip analog classifier, machine learning, late fusion, electrocardiogram

I. Introduction

Sepsis is a leading cause of death worldwide, and 80% of patients have sepsis onset outside hospital settings. Real-time, at-home health monitoring for at-risk patients is a potential solution for predicting sepsis onset and providing timely intervention. There are several challenges and limitations associated with the current technologies for at-home monitoring – 1. existing wearable devices lack the ability to integrate electronic medical record (EMR) with sensor data in real-time; 2. most wearable devices do not have automated inference capability and depend on telemetry and medical experts for actionable inference; 3. transmission of patient data over the network increases risk of breaches [1], [2]. The proposed work addresses these challenges through a two-step fusion AI framework - an AI circuit that can be integrated with wearable device for performing in-situ analysis of continuous sensor data, and a cloud AI model that performs fusion of demographic data and scores from embedded AI circuit for real-time risk prediction of sepsis onset. Raw patient data collected through wearable device is not transmitted to the cloud; rather only prediction scores of the embedded AI circuit is sent to the cloud for fusion which improves robustness of patient data.

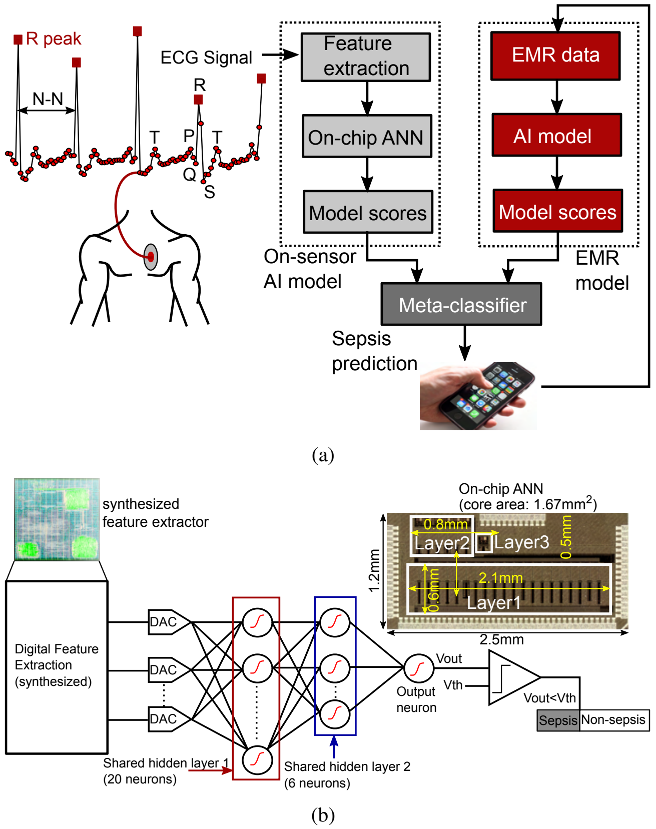
The proposed fusion AI framework has 3 components 1) analog on-chip machine-learning classifier that analyzes continuous ECG signal and predicts risk of sepsis; 2) random forest classifier based EMR model that predicts sepsis risk from demographics (age, gender, race and ethnicity) and co-morbidity data; (3) a meta-classifier that combines prediction scores from on-chip classifier and EMR model for prediction of sepsis onset (see Fig. 1a)). We have developed a mobile application that allows users to input their demographic and co-morbidity information, and predicts risk of sepsis onset. The on-chip classifier can be embedded into a wearable sensor and comprises of a fully integrated, 3-layer artificial neural network (ANN) for predicting sepsis from ECG sensor signal as shown in Fig. 1b). The ANN uses switched-capacitor (SC) compute-in-memory (CIM) followed by analog activation circuits. Compared to SRAM-CIM, SC-CIM computes vector matrix multiplications with higher linearity at the cost of re-configurability of on-chip AI model weights. The proposed fusion framework is demonstrated on patient data collected from Emory University Hospital over 2014-2018.

II. ANN design for Sepsis prediction from ECG

A. Dataset

With the approval of Emory Institutional Review Board (IRB), de-identified sepsis dataset is obtained from Emory University Hospital (EUH). Since we used only de-identified data and no patient communication has been made during the study, need of informed consent is waived by Emory IRB. The cohort consists of 965 patients admitted to ICUs at two hospitals within the Emory Healthcare system in Atlanta from 2014 to 2018. For each patient, there is at least 8 hours of ECG signal recordings from the time of admission in ICU, with the ECG signals sampled at 300Hz. Table I presents the demographics for both sepsis positive and negative patients. Our cohort mostly consist of patients older than 56 years.

The Third International Consensus Definition of Sepsis (Sepsis-3) [3], criterion was used to assign sepsis onset time (tsepsis-3) when two conditions were simultaneously satisfied: 1) there was a clinical suspicion of infection and; 2) there was a 2-point increase in SOFA score (tSOFA). According



(a)

(b)

Fig. 1: a) Overview of proposed real-time sepsis detection using fusion AI model that combines physiological and EMR data b) 3-layer ANN for prediction of sepsis from ECG signal

TABLE I: Patient characteristics table

Characteristics		***Summary***	
		Sepsis	***Non-sepsis***
Data		514 (53.26%)	451 (46.73%)
Gender	Male	281 (29.12%)	214 (22.17%)
	Female	233 (24.14%)	237 (24.56%)
Race	African American	191 (19.79%)	190 (19.69%)
	Caucasian/White	278 (28.8%)	215 (22.28%)
	Asian	8 (0.8%)	7 (0.7%)
	Hispanic	1 (0.1%)	0 (0%)
	Multiple	1 (0.1%)	2 (0.2%)
	American Indian/ Alaskan Native	0 (0%)	2 (0.2%)
	Unknown	35 (3%)	35 (3%)
Ethnicity	Hispanic/Latino	9 (0.9%)	7 (0.7%)
	Non-Hispanic/Latino	413 (42.78%)	357 (36.99%)
	Unknown	92 (9.53%)	87 (9.01%)
Age	16-35 years	52 (5.39%)	42 (4.35%)
	36-55 years	162 (16.79%)	149 (15.44%)
	56 and above years	300 (31.08%)	260 (26.94%)

to Sepsis-3 definition, 514 patients met the Sepsis-3 criterion. The dataset is randomly partitioned into 885 training samples and 80 test samples.

B. Feature extraction

The digital feature extractor (FE) in Fig. 1b) computes features on 30 second windows of the ECG signal. Only time-domain features are used in this work for low-cost implementation. 14 time-domain features are calculated from first-order statistical measures of location and distribution of R peaks, QRS complexes, PR intervals, ST intervals, QT intervals, and NN intervals (see Fig. 1). The FE removes baseline wander by subtracting median value from each segment. The digital features are converted into analog voltages using 4-bit SC DACs which drive the analog ANN.

C. ANN model training and circuit design

The ANN has 20 neurons in the first hidden layer, and 6 neurons in the second hidden layer (Fig. 1(b)). The hidden layers use custom tanh activation function, while the output layer uses a custom softmax activation function. The voltage output of the softmax function is compared with a threshold voltage (Vth) to generate the ANN decision, i.e., non-sepsis/sepsis. The activation circuits are designed using single-stage, common-source differential amplifiers as shown in Fig. 2. The fully differential amplifiers in the hidden layers use output offset cancellation technique to reduce amplifier offset. Offset in the output layer is removed through foreground calibration as described later.

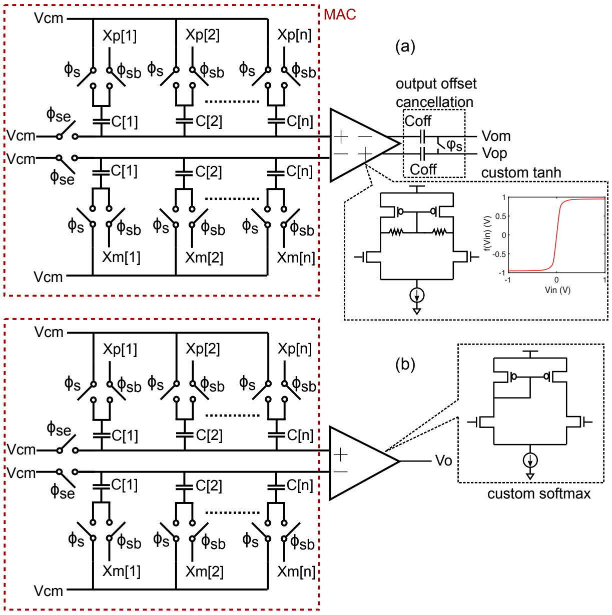


Fig. 2: Circuit schematic of a) hidden neuron with custom tanh activation b) output neuron with custom softmax neuron

The custom analog activation functions resemble their ideal, mathematical counterparts, but are not exactly the same. To ensure good matching between software ANN model and IC measurements, we use a hardware-software co-design methodology in which amplifier transfer curves, and their derivatives, are used to train the ANN model iteratively [4]. Stochastic gradient descent is used to optimize the ANN model by minimizing the loss function at each epoch. Once the ANN is fully trained, the model weights are encoded as capacitor values in the SC-CIM. The ANN weights are quantized to 4-b in the hidden layers, and 6-b in the output layer. Weight quantization is done during the training iterations to minimize effect of quantization error. A 4fF unit capacitor is used to realize an LSB weight in the SC-CIM. The unit capacitor value is selected to ensure capacitance mismatch does not degrade ANN accuracy.

III. EMR MODEL

In addition to ECG signal data, we also incorporated the patient demographics and basic co-morbidities. For demographics, we considered four data elements - age, race, gender and ethnicity. Prior co-morbidities of the patients are coded as International Statistical Classification of Diseases and Related Health Problems (ICD) 10th revision. In order to maintain standardization, we created mapping between disease names and the nearest ICD-10 codes. The EMR data underwent a series of pre-processing steps prior to formal analysis and model development. Given the wide range of ICD-10 codes (70,000 codes), we leverage the ICD-10 disease description as string to obtained the embeddings of multiple comorbidities. The EMR data elements were both categorical (gender, age bins) and textual (ICD-10 code descriptions) datatype. As the first data pre-processing step, we applied standard data cleaning steps, including removing empty cells and special characters. For conversion of categorical features to numerical quantities, we use the label encoding technique that converts each value in a column to a specific number.

The vectorization of ICD-10 code descriptions was performed using term frequency- inverse document frequency (Tf-idf) algorithm. Typically, the Tf-idf weight is composed of two terms: the first term computes the normalized term Frequency (TF) which is given by TF(n) = (number of times term n appears in a document) / (total number of terms in the document). The second term in the TF-idf weight is the inverse document frequency (IDF) which is computed as the logarithm of the number of documents in the corpus divided by the number of documents where the specific term appears. We trained the Tf-idf tokenizer using our training dataset and obtained 965×20 dimensional vector representation of the co-morbidities. Finally the numeric representation of the categorical features and Tf-idf representation of the co-morbidities are combined using linear concatenation. We standardized features by removing the mean and scaling to unit variance. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. The mean and standard deviation values are then used on later data (i.e, holdout test) using the same transformation function. In order to perform a comprehensive analysis, we experimented with multiple parametric (logistic regression, ANN) and non-parametric machine learning models (linear support vector machine - SVM, Random Forest). Given the static nature of this data, temporal sepsis onset prediction is not relevant with EMR and we only design a single-point prediction model for distinguishing sepsis versus non-sepsis data points using only EMR data. The optimal value of the hyper-parameters is tuned through 10-fold cross validation on the training data. Table II shows the accuracy of the different EMR models on the test set. Random forest achieves the highest accuracy and is used for fusion with predictions from sensor data.

TABLE II: Sepsis prediction results with EMR models

	Linear SVM	**Logistic regression**	**Random forest**	**ANN**
Accuracy (%)	49	53	76	51

IV. MEASUREMENT RESULTS

The ANN is fabricated in 65nm CMOS process, and has a core area of 1.67mm2^2 (Fig. 1(b)). The FE and DACs are implemented off-chip. Fig. 3(a)-(b) show the measured confusion matrices with on-chip ANN and after fusion respectively, on the test-set. Offset in the amplifier in the output layer of the ANN is calibrated by applying the training samples to the test-chip and setting the threshold voltage, Vth, to maximize prediction accuracy on training samples. The on-chip ANN detects sepsis with 85% accuracy from only ECG signal 4 hours before onset, while the accuracy improves to 91% after fusion with demographics and co-morbidity data. Linear support vector machine (SVM) is used as meta-classifier for fusion. Fig. 3c) shows measured accuracy versus time before sepsis onset (tonset). Accuracy of sepsis prediction from ECG signal increases closer to actual onset. The on-chip ANN consumes 7.1μW/inference at 1kHz operating frequency from 1.1V supply, while the DAC consumes 3.8μW (Fig. 3d)). The FE is digitally synthesized and consumes 2.1μW. Thus, the complete on-chip AI circuit has an estimated energy consumption of 12.9nJ/inference. The power consumption will increase to 13.6μW if analog front-end amplifier and 14-bit ADC for digitizing ECG signal is integrated on-chip. Fig. 4a) plots the measured accuracy as a function of supply voltage. The ANN accuracy reduces from 90% to 83.8% as the supply voltage is scaled from 1.2V to 0.8V, while the accuracy after fusion reduces from 92.5% to 88.8%. Fig. 4b) plots the measured accuracy for 5 test-chips. The mean accuracies of the on-chip ANN and fusion model are 86.5% and 92.2% respectively. Each test-chip is calibrated to suppress offset in the output layer. Fig. 4c)-d) plot the measured histogram of accuracy of the ANN and fusion model for 500 repeated evaluations. The low standard deviation in accuracy demonstrates relative robustness against noise. Figure 6 compares this work with state-of-the-art works. The proposed fusion framework achieves the highest accuracy while using single modality sensor data source and no laboratory test results, which demonstrates the feasibility of the proposed technique for at-home monitoring and is a key differentiation from state-of-the-art as shown in Table III. To the best of our knowledge, no works have demonstrated on-chip solution for sepsis prediction. Compared to a fully digital baseline model synthesized in 65nm, the proposed implementation achieves similar accuracy and almost 4.5× lower energy. The proposed classifier achieves 4x lower energy than state-of-the-art machine learning ASICs for different biomedical applications as shown in Fig. 5 thanks to analog implementation of the ANN.

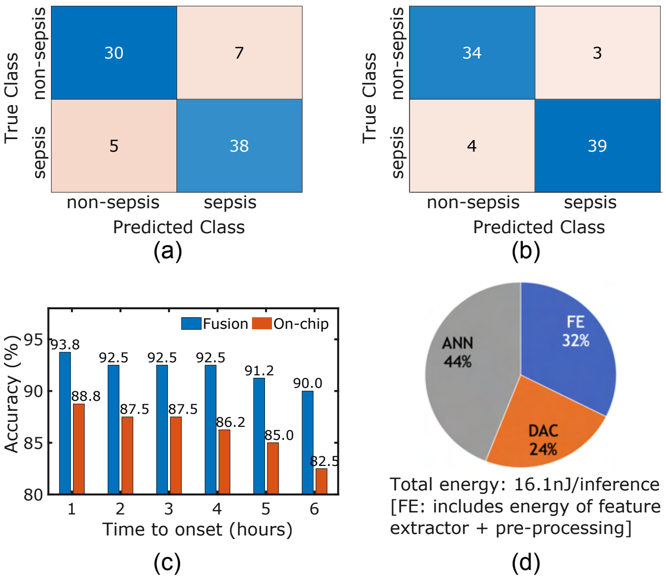


Fig. 3: Measured confusion matrices for a) on-chip ANN b) fusion model; c) accuracy vs time to onset d) energy breakdown

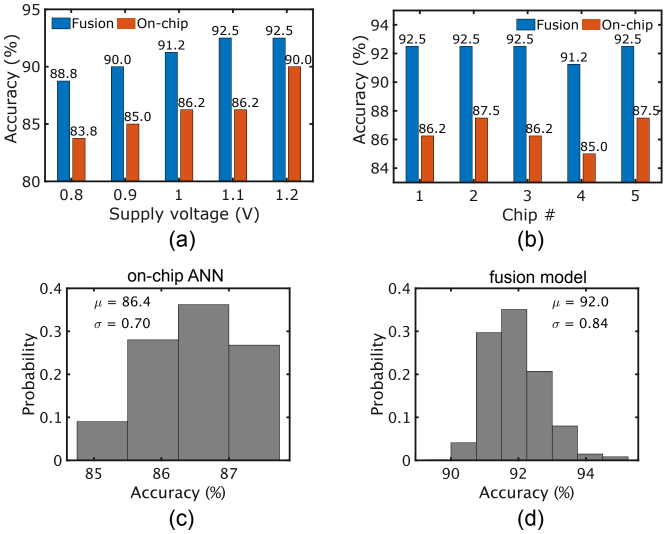


Fig. 4: Measured performance vs a) supply voltage b) multiple chips; repeated evaluations with c) on-chip ANN d) fusion model

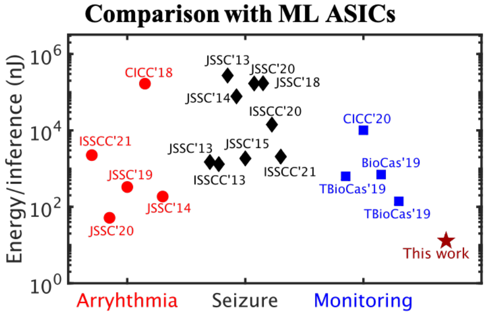
V. Conclusion

This work has presented a fusion AI framework for sepsis prediction 4 hours before onset. In our future work, we will collect multi-institutional data since data collected from a single hospital may have bias towards training population which is dominated by elderly and african-american population for this dataset. We will also perform a race, gender, ethnicity and age based disparity analysis to understand the effect of bias on various sub-groups.

TABLE III: Comparison with state-of-the-art software AI models

	Model	Performance Metrics			Data Source		
		t_{onset} (hrs)	Accuracy	Sensitivity	Vitals	Lab[1]	Dem.[2]
[5]	LSTM	4	84.5%	–	8	26	6
[6]	Ensemble	3	82.7%	0.81	9	0	0
[7]	Random forest	6	74.6%	–	8	26	6
[8]	Regression	6	86.4%	0.30	8	26	6
[9]	Survival model	4	67%	0.85	16	30	19
[10]	RNN	4	–	0.84	9	39	36
[11]	Random forest	4	–	0.87	5	6	4
[12]	GRU	6	99.8%	0.94	8	26	0
[13]	Regression	4	61%	0.55	8	0	7
[14]	LSTM	3	93%	0.94	8	0	1
[15]	LSTM+CNN	3	91.5%	0.97	6	37	35
This work	**On-chip model**	**4**	**86.5%**[3]	**0.80**[3]	**1**	**0**	**4**
	Fusion model	**4**	**92.2%**[3]	**0.92**[4]	**1**	**0**	**5**

[1] includes laboratory test results and culture results; [2] includes demographics and co-morbidities; [3] average of 5 test-chips



Comparison with digital baseline

	On-chip ANN	Digital baseline
Accuracy	86.5%[1]	87.5%
Sensitivity	0.80[1]	0.85
Energy/inference	16.1nJ[2]	72.2nJ[3]

[1] average of 5 test-chips; [2] includes energy of ANN, DAC and FE; [3] includes energy of FE and digital ANN

Fig. 5: Comparison with state-of-the-art ML ASICs and digital baseline

VI. Acknowledgements

This material is based on research sponsored by Air Force Research Laboratory under agreement number FA8650-18-2-5402. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

References

[1] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Security and privacy in the medical internet of things: a review," *Security and Communication Networks*, vol. 2018, 2018.

[2] J. Liu and W. Sun, "Smart attacks against intelligent wearables in people-centric internet of things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 44–49, 2016.

[3] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[4] S. T. Chandrasekaran, A. Jayaraj, V. E. G. Karnam, I. Banerjee, and A. Sanyal, "Fully Integrated Analog Machine Learning Classifier Using Custom Activation Function for Low Resolution Image Classification," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 3, pp. 1023–1033, 2021.

[5] B. Roussel, J. Behar, and J. Oster, "A Recurrent Neural Network for the Prediction of Vital Sign Evolution and Sepsis in ICU," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[6] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, and R. Das, "A computational approach to early sepsis detection," *Computers in biology and medicine*, vol. 74, pp. 69–73, 2016.

[7] M. Nakhashi, A. Toffy, P. Achuth, L. Palanichamy, and C. Vikas, "Early Prediction of Sepsis: Using State-of-the-art Machine Learning Techniques on Vital Sign Inputs," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[8] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, and T. Lyons, "The signature-based model for early detection of sepsis from electronic health records in the intensive care unit," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[9] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the ICU," *Critical care medicine*, vol. 46, no. 4, p. 547, 2018.

[10] A. D. Bedoya, J. Futoma, M. E. Clement, K. Corey, N. Brajer, A. Lin, M. G. Simons, M. Gao, M. Nichols, S. Balu *et al.*, "Machine learning for early detection of sepsis: an internal and temporal validation study," *JAMIA open*, vol. 3, no. 2, pp. 252–260, 2020.

[11] K. H. Goh, L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[12] S. D. Wickramaratne and M. S. Mahmud, "Bi-Directional Gated Recurrent Unit Based Ensemble Model for the Early Detection of Sepsis," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 70–73.

[13] S. P. Shashikumar, M. D. Stanley, I. Sadiq, Q. Li, A. Holder, G. D. Clifford, and S. Nemati, "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics," *Journal of electrocardiology*, vol. 50, no. 6, pp. 739–743, 2017.

[14] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Computers in biology and medicine*, vol. 89, pp. 248–255, 2017.

[15] C. Lin, Y. Zhang, J. Ivy, M. Capan, R. Arnold, J. M. Huddleston, and M. Chi, "Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2018, pp. 219–228.