



International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673

<https://www.inderscience.com/ijdmb>

Dual pipeline technique for detecting sepsis from photoplethysmography

Shadi Abudalfa, Sara Lombardi, Eleonora Barcali, Leonardo Bocchi

DOI: [10.1504/IJDMB.2025.10066271](https://doi.org/10.1504/IJDMB.2025.10066271)

Article History:

Received:	26 November 2023
Last revised:	16 February 2024
Accepted:	26 June 2024
Published online:	10 July 2025

Dual pipeline technique for detecting sepsis from photoplethysmography

Shadi Abudalfa*

SDAIA-KFUPM Joint Research Center for Artificial Intelligence,
King Fahd University of Petroleum and Minerals,
Dhahran, 31261, Kingdom of Saudi Arabia
Email: shadi.abudalfa@kfupm.edu.sa

*Corresponding author

Sara Lombardi, Eleonora Barcali and
Leonardo Bocchi

Department of Information Engineering,
University of Florence,
Florence, 50139, Italy
Email: sara.lombardi2@unifi.it
Email: eleonora.barcali@unifi.it
Email: leonardo.bocchi@unifi.it

Abstract: The goal of this work is to improve the performance of sepsis-detection in photoplethysmography (PPG) data. To achieve this goal, we present a hybrid technique for classifying sepsis in PPG data based on confident learning (CL) with noisy data. The technique presented in this study employs CL to improve the accuracy and reliability of the machine learning models, as it takes into account the uncertainty associated with each prediction. Numerous experiments were carried out to assess the performance of the presented technique in detecting sepsis using PPG data. The results obtained, using the best-performing XGBoost model, were compared with those of a previous study in which a deep learning-based model was applied to the same sample of data. The presented technique demonstrated its effectiveness by achieving an F1-score of 80.62% on test set, with a 7% improvement compared to the performance of the previous study.

Keywords: confident learning; rich features; noisy data; photoplethysmography; sepsis; synthetic data generation.

Reference to this paper should be made as follows: Abudalfa, S., Lombardi, S., Barcali, E. and Bocchi, L. (2025) 'Dual pipeline technique for detecting sepsis from photoplethysmography', *Int. J. Data Mining and Bioinformatics*, Vol. 29, No. 3, pp.362–380.

Biographical notes: Shadi Abudalfa received his BSc and MSc both in Computer Engineering from the Islamic University of Gaza (IUG), Palestine in 2003 and 2010 respectively. In 2013, he joined King Fahd University of Petroleum and Minerals (KFUPM) as a full time student (scholarship) to pursue his PhD. In 2018, he completed his PhD in Computer Science

and Engineering at KFUPM. He has a strong teaching experience though his work as an Assistant Professor at the University College of Applied Sciences, Palestine. He is currently a researcher at SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI). From July 2003 to August 2004, he worked as a research assistant at Projects and Research Lab in IUG. During same period, he worked as a teaching assistant at Faculty of Engineering in IUG. Abudalfa serves as reviewer through various conferences and journals. His current research interests include artificial intelligence, data mining, activity recognition.

Sara Lombardi is a Biomedical Engineer and currently a PhD student at the Department of Information Engineering of the University of Florence. Her academic background includes a Bachelor's in Electronic and Telecommunications Engineering and a Master's in Biomedical Engineering. In 2022, she worked as a Research Fellow at the Department of Information Engineering, University of Florence, in the area of physiological signal modelling for microcirculation analysis. Her research interests include biosignal analysis, bioimage processing and artificial intelligence. She is a member of the IEEE and the Engineering in Medicine and Biology Society.

Eleonora Barcali is a Biomedical Engineer registered with the Order of Engineers of Florence, a Doctoral student in the Department of Information Engineering at the University of Florence, and currently a Visiting Doctoral student at the Université Libre de Bruxelles. Her academic background includes a Bachelor's in Mechanical Engineering and a Master's in Biomedical Engineering. Her doctoral project concerns the application of radiomics in medicine to support radiologists in performing quantitative and objective diagnosis and prognosis in an innovative way. She wants to apply classical radiomics primarily to bone using both artificial intelligence and biomechanics. Her research interests include biomedical signal processing, artificial intelligence, biomechanics and radiomics in imaging. She is a member of the Italian National Bioengineering Group.

Leonardo Bocchi is an Associate Professor of Biomedical Engineering at the Department of Information Engineering, University of Florence. Information Engineering at the University of Florence. His research interests include assisted diagnosis tools based on image and medical signal processing, artificial intelligence techniques based on soft computing, and modelling of physiological systems and cognitive processes. Major applications include the detection and classification of breast tumours based on ultrasound, X-ray, or MRI images; the study of microcirculatory dynamics and their relationships with chronic or complex pathological conditions such as diabetes or sepsis; and cognitive and behavioural modelling using non-intrusive monitoring techniques such as motion analysis, facial expression properties, and sentiment analysis. He is a member of the IEEE-Engineering in Medicine and Biology Society and a member of the IEEE-EMBS Technical Committee on Cardiopulmonary Systems.

1 Introduction

Sepsis (Singer et al., 2016) is a critical condition that poses a significant threat to patients in ICUs (Sakr et al., 2018) and remains one of the primary causes of death (Rudd et al., 2020). This dreaded clinical condition is characterised by microcirculatory dysfunctions that results in reduced tissue perfusion and strongly correlates with the development of organ damage and increased mortality (De Backer, 2023; Charlton et al., 2017). A dysfunction of peripheral tissue perfusion can be evaluated at the skin level by using the photoplethysmogram (PPG).

PPG is an optical signal that shows blood volume changes and oxygen saturation at a peripheral site, according to the absorption or reflection of light through the blood. This signal can be obtained with the pulse oximeter, a noninvasive device that is commonly used in hospital settings for cardiovascular monitoring of patients.

PPG is therefore an effective measure for monitoring sepsis-related anomalies. This has encouraged researchers to explore numerous techniques that can analyze PPG signals and detect such microcirculatory abnormalities. One research direction is based on employing machine learning techniques for detecting sepsis in PPG data (Fleuren et al., 2020). Whereas, there are still many challenges that need to be addressed to improve the performance of sepsis identification in PPG data. In this work, we present a technique to detect sepsis with competitive performance. The purpose is to solve the problem of detecting sepsis using just PPG data. The contributions of this work are highlighted below:

- We designed a hybrid technique by combining two confident learning (CL) models based on gradient boosting (GB) of regression trees. According to our knowledge, our research is the first one that employs this hybrid technique for detecting sepsis from PPG signals.
- The presented technique is employed for developing an integrated system that detects sepsis from PPG signals, which could be applied in ICUs.
- We evaluated performance of the presented technique by using a public database and reported results through various perspectives.
- The presented technique improves the performance of classifying sepsis from PPG data by manipulating the added noise. This technique mimics diffusion models (Yang et al., 2023) for generating data akin to the data used to train the GB model.

This work is an extension to a work achieved by Lombardi et al. (2022b). In the previous study, they had developed a deep learning-based model with the aim of identifying sepsis from the analysis of the PPG signal only. In this work, we start from the same dataset provided in the previous study and evaluate the performance of our new technique with the aim to improve the resulting performance.

The paper is structured as follows: a review of related studies is presented in Section 2, followed by an explanation of the proposed technique in Section 3. Section 4 details the experimental environment, while Section 5 discusses the experimental results and provides the analysis. The paper is concluded in Section 6, which also includes suggestions for future work.

2 Literature review

Sepsis is a rapidly progressive clinical condition for which the mortality rate is connected to the timeliness of therapeutic intervention. International guidelines (Evans et al., 2021) emphasise that early identification and appropriate management in the early hours of sepsis development are of paramount importance to improve patient outcomes. For this reason, many studies focus on finding new automated methods for early sepsis identification.

In clinical practice, several variables and tabulated scoring systems are used for sepsis screening such as the systemic inflammatory response syndrome (SIRS) criteria (Rangel-Frausto et al., 1995), the sequential organ failure assessment (SOFA) (Vincent et al., 1996) and the quick sequential organ failure score (qSOFA) criteria (Singer et al., 2016). These tools assess various parameters such as fever, heart rate and blood pressure, in addition to indicators obtained from laboratory tests. These parameters, along with other variables have been used in several studies to develop machine learning approaches for sepsis identification.

Calvert et al. (2016) developed 'InSight', a machine learning algorithm that leveraged vital signs and demographic features to identify sepsis. Further studies (Desautels et al., 2016) validated the same algorithm by using different datasets and input features for demonstrating its superior performance compared to standard sepsis diagnostic methods such as tabulated scoring systems. Used features included variables such as age, systolic blood pressure, pulse pressure, heart rate, respiratory rate, SpO₂ and Glasgow Coma Score.

Mollura et al. (2020) diligently conducted extensive training sessions for a variety of machine learning classifiers, all with the primary goal of identifying sepsis cases within the crucial first hour of patients' admission to the intensive care unit (ICU). To accomplish this ambitious goal, the machine learning classifiers were furnished with essential features extracted from the continuously recorded signals of two critical sources: the electrocardiogram (ECG) and arterial blood pressure (ABP). These vital signals were used to empower the classifiers, enabling them to make accurate and timely assessments for the early detection of sepsis cases in ICU patients. The results presented by the authors demonstrated that the classification accuracy achieved by these machine learning models was similar to the scores obtained from conventional tabulated methods. This exciting finding suggests that the waveforms of vital signs could serve as valuable indicators for the early detection of sepsis.

Some studies have explored deep learning approaches based on the use of artificial neural networks. For example, Kam and Kim (2017) used periodic measurements of vital signs and laboratory parameters to train a long-short-term memory network (LSTM). Aşuroğlu and Oğul (2021) used a combination of random forest algorithm and convolutional neural network (CNN) to predict SOFA score using measurements of heart rate, blood pressure, respiratory rate, oxygen saturation, glasgow coma score eye opening and temperature. Similarly, Scherpf et al. (2019) developed a recurrent neural network (RNN) to predict sepsis onset using one hour-averaged values of blood pressure, heart rate, body temperature, respiration rate, white blood cell count, pH, blood oxygen saturation and age.

It is worth mentioning that most previous research studies used multiple clinical parameters, including blood test parameters and other indices commonly used for sepsis diagnosis, such as white blood cell count, to perform automatic sepsis detection.

However, some of these parameters are not immediately available at the patient's bedside, so they make the application of the method not suitable for a prompt screening procedure. Therefore, in our study, we focused our analysis only on the PPG time series, since this evaluation can be performed promptly and noninvasively by applying the saturimeter to the patient's finger.

The objective of this study is to investigate the feasibility of detecting sepsis-induced microcirculatory changes from PPG data. Through the review of the existing literature, we confidently assert that our work makes a significant contribution to the field of knowledge by bridging an important gap in the existing research.

3 Proposed architecture

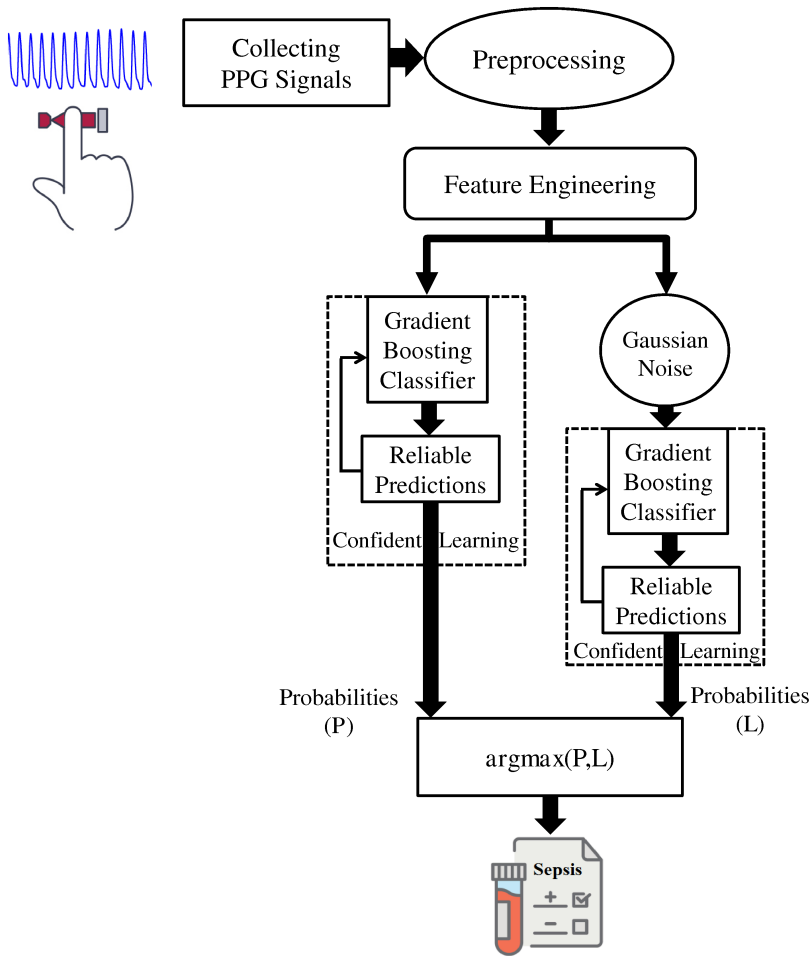
Our presented technique consists of several steps. The first one is based on collecting PPG signals from patients. Then, the preprocessing phase can be applied to PPG for decreasing effect of noise that may be appeared during its measurements (Heo et al., 2021). After that, a feature extraction module computes a rich set of features from the raw data. We extract features from the raw data for improving performance of detecting sepsis in PPG data.

Finally, we apply the classification task to the data extracted from the step of feature extraction for classifying each PPG signal into either sepsis or non-sepsis. For achieving accurate classification, we use cascading methods through two parallel pipelines. Figure 1 illustrates the structure of our presented technique. Whereas, the milestones of the two parallel pipelines are described in the following paragraphs.

Two machine learning models based on CL (Northcutt et al., 2021) are trained to perform the classification task through the two parallel pipelines. Each model is trained by using GB of regression trees (Friedman, 2001). The first model is immediately applied to the data resulted from the feature extraction step. While, the second one is applied to the data obtained by adding random Gaussian noise (GN) to the data extracted from the step of feature extraction. The outputs of the two parallel pipelines are combined using an ensemble method. This ensemble method is based on identifying the output label that trends to the maximum value of prediction probabilities provided by the two models. We employ GB classifier for their competitive performance relative to alternative classification methods.

CL is a machine learning approach that seeks to improve the accuracy and robustness of machine learning models by incorporating the notion of confidence in the learning process. In CL, the model is trained not only to make predictions, but also to provide an estimation of their uncertainty and confidence. This is achieved by assigning a confidence score to each model output. This process is iteratively continued until finding the best model with reliable predictions.

By incorporating confidence scores into the learning process, CL can help to identify and handle cases where the model is uncertain or likely to make inaccurate predictions. This can improve the overall accuracy and robustness of the model, especially in scenarios where the input data is noisy or uncertain. We use CL in this work for decreasing the effect of noise that is generated during the data collection and preprocessing phases. CL identifies and corrects mislabelled data points in a supervised learning setting. Thereby, the performance of GB of regression trees is improved after training it with CL.

Figure 1 The presented technique (see online version for colours)

4 Experimental setup

We conducted many experiments to evaluate the performance of the presented technique for classifying sepsis from PPG signals. All experiments were conducted by using Python language within the Google Colaboratory environment (Bisong et al., 2019). It is worth mentioning here that we used only CPU processor for conducting all our experiments. Table 1 shows the Python modules employed in the development of our technique.

4.1 Dataset

In our study, we used the same dataset built and described by Lombardi et al. (2022b). In this subsection, we decided to report a quick overview of the used dataset to show its characteristics. The dataset comprised a set of PPG samples extracted using the popular

MIMIC-III database (Johnson et al., 2016) along with the MIMIC waveform database matched subset (Moody et al., 2020). The first database contains clinical data collected from subjects admitted to the ICU, while the second one contains the corresponding waveforms acquired by multi-parameter monitor during the patient hospitalisation.

Table 1 Python modules

<i>Tool</i>	<i>Version</i>	<i>Purpose</i>
numpy	1.25.2	Reading signal data and extracting traditional features
sklearn	1.2.2	Building and evaluating models
eeglib	0.4.1.1	Extracting features from EEG signals
tsfel	0.1.6	Feature extraction on time series data
cleanlab	2.4.0	Applying confident learning with noisy label

The dataset was obtained from the MIMIC III database, which includes clinical data from more than 40,000 patients admitted to the ICU in a medical centre. From this database, the authors selected two groups of subjects based on the ICD-9 diagnosis code with the aim of forming two groups. The first group, named sepsis, includes 85 patients with sepsis with varying degrees of severity (sepsis, severe sepsis, septic shock). The second group, named controls, contains 101 patients with five different clinical conditions other than sepsis [ICD-9: 311 (depressive disorder NEC), 3051 (tobacco use disorder), 30,000 (anxiety state NOS), 2,948 (other persistent mental disorders due to conditions classified elsewhere), 3,004 (dysthymic disorder)]. The PPG signal recordings related to the selected patients were derived from the MIMIC waveform database matched subset, which contains waveforms recorded by multi-parameter monitors during ICU stay.

The PPG recordings for the selected subjects were then processed to derive a set of fixed-length PPG samples of good quality. To achieve this goal, the authors implemented a specific template-matching algorithm, in which they evaluated waveform morphology (Lombardi et al., 2022a). Specifically, after removing missing values from each collected waveform, the signal acquired from the patient was compared with a synthetic ideal PPG signal by using Pearson’s correlation coefficient. The introduction of a minimum threshold for the correlation coefficient, equal to 0.8, made it possible to select only a set of PPG samples that were highly correlated with the reference signal and therefore of good quality.

Subsequently, the authors chose a fixed length for the PPG segments. The chosen duration was 15,000 samples corresponding to two minutes of acquisition duration. The implementation details of the preprocessing algorithm are described in detail in the article by Lombardi et al. (2022a).

The data is divided into two classes, called ‘sepsis’ and ‘control’ based on ICD-9 diagnosis codes. The sepsis group contains PPG signals related to 85 subjects affected by sepsis with varying degrees of severity, while the control group contains signals acquired in 101 subjects with a subset of five ICD-9 diagnoses different from sepsis. Each signal is previously preprocessed with the aim of deriving a set of good-quality fixed-length PPG segments, as detailed in a previous study presented by Lombardi et al. (2022b).

The processed data were divided by the authors into training set and test set by randomly assigning 80% of the subjects in each group to the training set (68 septic and

81 control subjects) and 20% to the test set (17 septic and 20 control subjects). Table 2 shows the number of samples located in each class of training and test sets. In this study we use the same data division provided by the authors so that we can compare their results with our approach. In the text, we refer to this data as raw data.

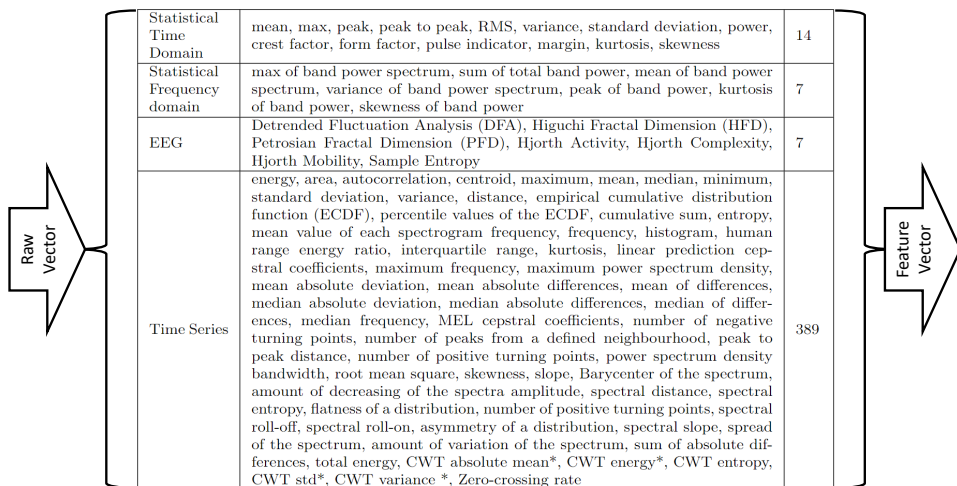
Table 2 Dataset description

Class type	Training set	Test set
Control group	6,600	1,598
Sepsis group	5,816	1,377
Total samples	12,416	2,975

4.2 Feature extraction

In this work, we extract a rich set of features from the raw data. There are three categories of features extracted from the raw data. The first set includes statistical features calculated by using Numpy module (Gambera, 2023). The second one contains features calculated by using Eeglib module (Cabanero-Gomez et al., 2021). This category extracts features that are usually taken from electroencephalographic (EEG) signals. We believe that some nonlinear features often used for the analysis of the EEG signal, such as entropy and fractal analysis, can also be studied in PPG signals.

Figure 2 Feature set



The third category of features is computed by using Tsfel module (Barandas et al., 2020). We use this set of features to mimic the extraction of features from PPG signals as it was a time series problem. As a result of this extraction step, each sample in the dataset includes 417 values.

Figure 2 shows all features extracted from the raw data. It is worth mentioning here that some features of time series provides more than one value since it is calculated for each wavelet scale. These characteristics are indicated by symbol * in the figure.

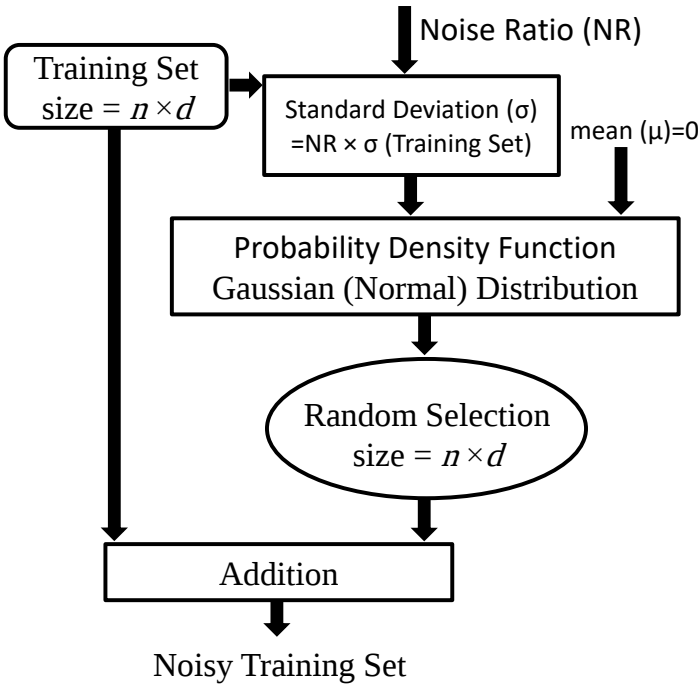
We used this set of features because it provides competitive performance with respect to the previous study conducted on the same dataset. Furthermore, we did not reduce the number of features used because, based on our experiments, applying a feature selection strategy does not improve the performance.

4.3 Generating GN

As stated in Section 3 one of the two pipelines in our technique is based on adding GN to the extracted features. In this work, we use a specific process to add noise to the training set data. This process gives the opportunity for training the model of CL with noisy data. Figure 3 illustrates a flowchart of the method used for adding GN to the training set.

The noisy data added to the training set is randomly selected from a Gaussian (normal) distribution. The probability density function used with Gaussian distribution is based on setting values of mean μ and standard deviation σ . The mean represents the central tendency of a normal distribution and determines the position of the peak in the bell curve (Devlin et al., 1997). The standard deviation serves as a metric for measuring variability within a dataset. It quantifies the width of the normal distribution and indicates how the values are spread out with respect to the mean.

Figure 3 Adding GN to the training set



In our experiments, we set the mean μ value to 0. Whereas, the value of standard deviation σ is firstly calculated by computing standard deviation for the training data. Then, the resultant value is multiplied by a factor named noise ratio (NR). The use of NR allows to control the level of noise added to the training set. It is worth mentioning that implementing this process led to data of the same size as the training set since the modification is accomplished by altering the feature values.

4.4 Evaluation measures

The empirical results obtained from the experiments provide a good way to evaluate the performance of the technique presented. In this study we used classification accuracy (Acc) and F1-score (Metz, 2008) as essential evaluation metrics. The accuracy represents the proportion of correctly classified samples in relation to the total number of samples. While, the F1-score is a significant measure that takes into account both precision and recall metrics. The F1-score is essentially the harmonic mean of precision and recall, with its optimal value being 1 and the worst value being 0.

To gain a more comprehensive understanding of F1-score, we delve deeper into the concepts of precision and recall. Recall, also known as sensitivity or true positive rate, quantifies the proportion of positive samples that were correctly classified as positive out of all the actual positive samples. On the other hand, precision denotes the proportion of positive samples that were correctly classified as positive, but this time, out of all the samples that were classified as positive by the model.

Additionally, to show the comparison with the previous study, we report the performance of the presented technique in terms of specificity and sensitivity. Sensitivity, which referred to as true positive rate, measures the ability to correctly identify samples with the disease. While, specificity, referred to as true negative rate, measures how well a classifier can identify true negatives samples. In this study, negative samples are included in the control group whereas positive samples are belong to the sepsis group.

5 Results and analysis

Several experiments were performed to assess the effectiveness of the presented technique. The experiments include the use of different machine learning models on both the raw dataset and the data obtained from the feature extraction process. In addition, the results obtained from the application of different configurations for CL and GN are presented.

5.1 Evaluating the raw data

Firstly, we evaluated the performance of numerous machine learning algorithms applied to the raw data described in Subsection 4.1. Based on our experiment work, the raw data poses challenges when we attempt to apply it to our classification problem. Table 3 shows all the results obtained from this experimental work. It is clear that the performance is limited when using the raw data. This result encouraged us to add a stage of feature extraction to our technique as illustrated in Section 3.

Table 3 Applying machine learning algorithms to the raw data

<i>Model</i>	<i>Accuracy</i>	<i>F1</i>	<i>Specificity</i>	<i>Sensitivity</i>
Histogram-based gradient boosting	60.91	66.44	59.66	61.65
Gradient boosting	59.87	65.35	58.12	60.93
Light GBM	59.87	64.99	57.87	61.15
XGBoost	58.39	62.35	55.41	60.65
Random forest	60.54	67.71	60.83	60.40
Logistic regression	50.15	55.48	45.73	53.32
Decision tree	52.24	55.50	48.41	55.55
K-nearest neighbours	56.10	45.58	51.63	68.20
Gaussian Naive Bayes	54.02	54.09	50.28	58.32
Support vector machine	57.65	65.44	56.31	58.25
AdaBoostClassifier	55.26	58.34	51.67	58.36

5.2 Evaluating the feature set

We evaluated the performance of applying the machine learning algorithms, previously used on the raw data, to the data resulting from the feature extraction phase, described in Subsection 4.2. The results provided by these experiment are reported in Table 4. We can clearly note that the performance of classifying PPG data is significantly increased.

Table 4 Applying machine learning algorithms to the feature set

<i>Model</i>	<i>Accuracy</i>	<i>F1</i>	<i>Specificity</i>	<i>Sensitivity</i>
Histogram-based gradient boosting	72.30	75.21	72.14	72.42
Gradient boosting	74.29	76.77	73.91	74.57
Light GBM	73.24	76.05	73.26	73.23
XGBoost	71.80	74.71	71.42	72.08
Random forest	71.03	73.48	69.46	72.28
Logistic regression	71.06	76.46	78.17	67.90
Decision tree	65.11	67.92	62.68	67.09
K-nearest neighbours	68.81	72.63	68.98	68.69
Gaussian Naive Bayes	54.02	54.09	50.28	58.32
Support vector machine	53.71	69.89	NaN	53.71
AdaBoostClassifier	72.94	75.55	72.34	73.39

It is worth mentioning that we evaluated other methods for extracting features. Whereas, the feature set described in Figure 2 are the ones that provide the best performance. We tried also to extract some features by using Neurokit2 (Makowski et al., 2021). This Python module is usually used for extracting features that are originally generated from PPG signals. Our experiments showed that using such features does not result in improved performance compared to the feature set described in Figure 2.

5.3 Performance of the presented technique

Our technique involves the use of two parallel pipelines that respectively include the application of GB and CL to the original feature set along with the application of GB and CL to the feature set resulting from the addition of GN. To show performance of the presented technique, we report the results obtained from each pipeline. In this experiment, we used GB algorithm as it yielded the best results when applied to the feature set (Table 4).

For the application of CL, we used the default settings which are detailed in Table 5. We only tuned the seed parameter for evaluating its impact on the performance. We used ten pseudo-random seeds for conducting this experiment. The used numbers are reported in order as follows: [3, 10, 42, 1,234, 5,678, 15,678, 125,678, 2,345,678, 12,345,678, 102,345,678]. Varying the seed parameter when applying GB and CL to the feature set produced a variability range of (74.43 ± 0.23) for accuracy, (77.09 ± 0.22) for F1-score, (75.13 ± 0.24) for specificity and (74.95 ± 0.11) for sensitivity. The boxplots shown in Figure 4 describe the distribution of evaluation metrics with a 95% confidence interval when using the ten values for the seed parameter tuning. We also repeated this experiment with ten seeds selected from the range [1, 10] to illustrate that similar finding is revealed with continuous range of seeds as shown in Figure 5.

Table 5 Setting used for CL

<i>Parameter</i>	<i>Description</i>	<i>Used value</i>
seed	Set the default state of the random number generator used for cross-validation	Different values
cv_n_folds	Number of cross-validation folds used to compute out-of-sample probabilities	5
converge_latent_estimates	If true, enforces numerical consistency of latent estimates	False
pulearning	Set the perfectly labelled class (works for two class datasets only)	None
find_label_issues_kwargs	Keyword arguments for filter.find_label_issues	None
label_quality_scores_kwargs	Keyword arguments for rank.get_label_quality_scores	None
verbose	Controls the verbosity of the output	False

Additionally, we evaluated performance of tuning the seed parameter when adding GN to the training set. In this experiment, we applied GB and CL to the data of features set with different ratios of added GN. We also tuned the seed parameter with setting NR value to 0.05, since this value provides local optimal results with the presented technique. Figure 6 shows confidence interval with confidence level equals to 95% when tuning the seed parameter after the application of GN. In this experiment we used the same ten pseudo-random seeds previously reported. The figure shows confidence intervals with respect to evaluation measures: accuracy with 76.83 ± 0.14 , F1 with 79.87 ± 0.12 , Specificity with 80.0 ± 0.52 and Sensitivity with 74.74 ± 0.22 . We also repeated this experiment with 10 seeds selected from the range [1, 10] to illustrate that similar finding is revealed with continuous range of seeds as shown in Figure 7.

Figure 4 Confidence interval with 95% confidence for applying CL with GB with arbitrary seeds (see online version for colours)

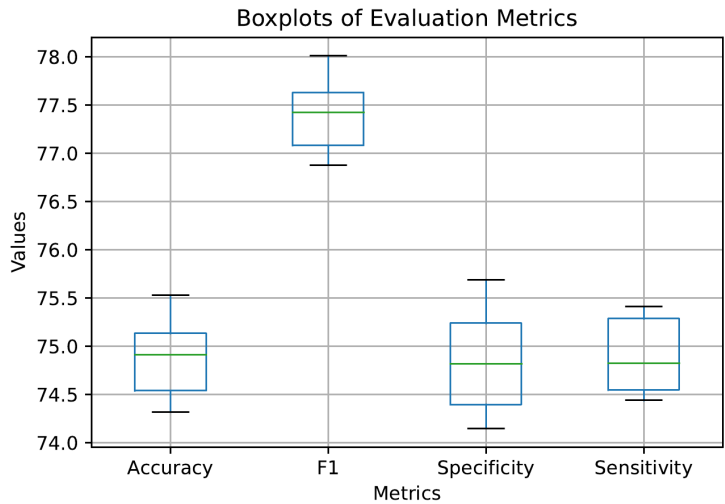
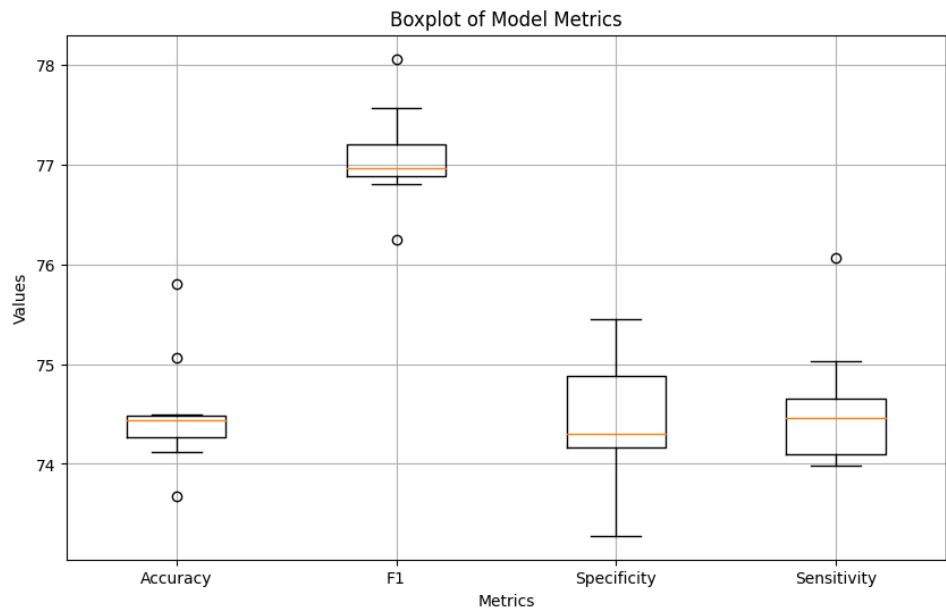


Figure 5 Confidence intervals with 95% confidence for applying CL with GB and GN for seeds from [1, 10] (see online version for colours)



We have summarised the results of our experiments in Table 6. This table allows the comparison of the different performances obtained by using each method mentioned in the paper. Here are the performance of the work from which we started, the machine learning models and the different methods attempted. In addition, it also makes clear the settings used for each experiment relative to the seeding point and NR parameters.

With ‘ours’ term in the last lines we refer to the application of all the steps of the entire technique presented with our work.

Figure 6 Confidence intervals with 95% confidence for applying CL with GB and GN with arbitrary seeds (see online version for colours)

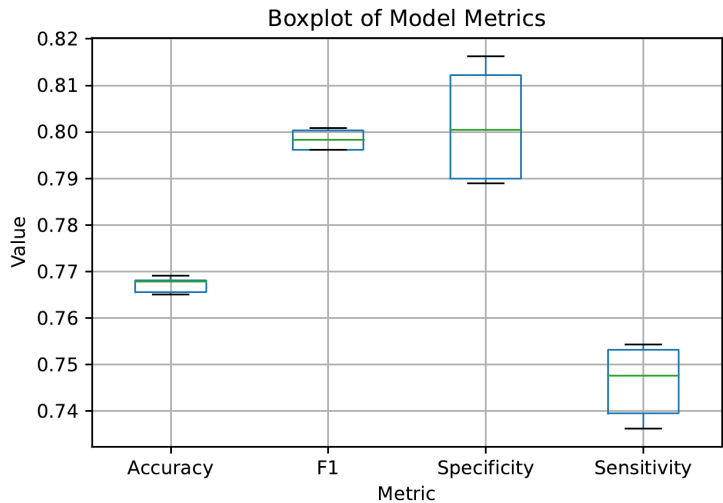
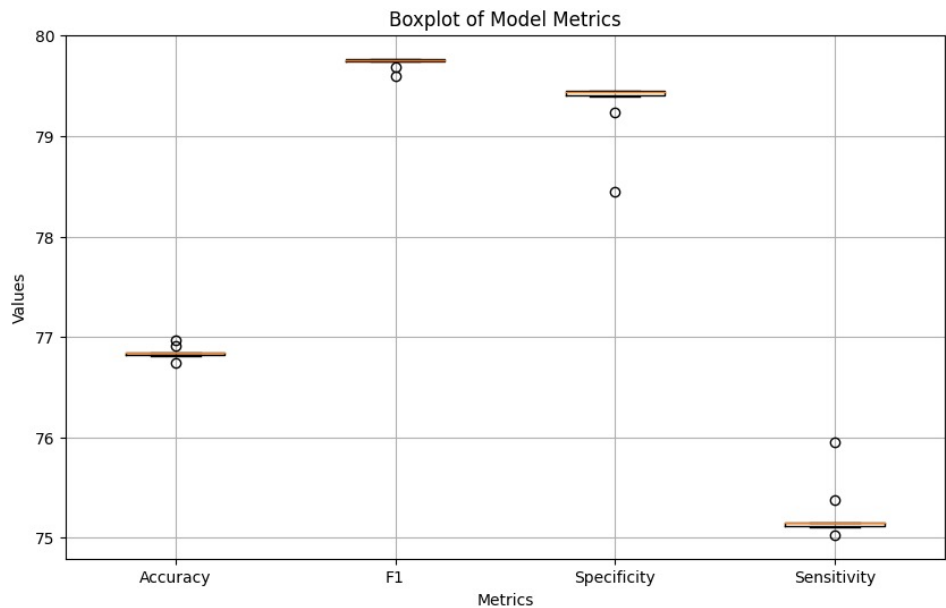


Figure 7 Confidence intervals with 95% confidence for applying CL with GB and GN for seeds from [1, 10] (see online version for colours)



Since our technique relies on the use of GN, described by NR parameter, we demonstrate the results obtained from the application of our technique with different values of the NR parameter. We firstly evaluated the performance of tuning NR parameter with values

starting from 0.001 to 0.01 with an increase step of 0.001. Then, we modified the range of values to be between 0.01 and 0.1 with an increase step of 0.01 followed by an increase step of 0.05 until reaching 0.45. Confidence intervals of results provided by this experiment are calculated and reported in Table 6 with a confidence level of 95%. Furthermore, Table 6 displays three optimal results yielded by this experiment.

Figure 8 Accuracy of the presented technique with modifying NR (see online version for colours)

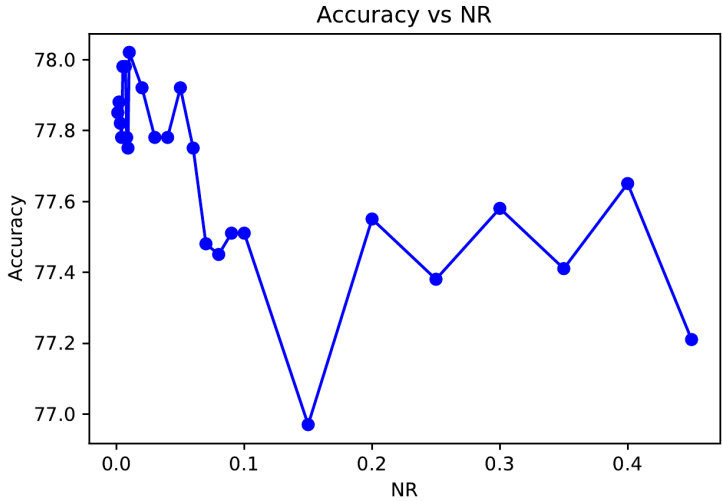


Table 6 Performance of the presented technique

Model	Settings	Acc	F1	Specificity	Sensitivity
Lombardi et al. (2022b)		76.37	73.54	81.04	70.95
ResNet		68.87	76.16	41.40	92.55
ZFNet		67.00	60.21	90.78	46.50
ShuffleNet		72.94	74.71	71.24	74.41
GB		74.29	76.77	73.91	74.57
CL + GB (mean)		74.43 ± 0.23	77.09 ± 0.22	75.13 ± 0.24	74.95 ± 0.11
CL + GB (optimal)	Seed = 1,234	75.53	78.01	75.69	75.41
GB + GN		76.13	79.89	81.98	72.98
CL + GB + GN (mean)		76.83 ± 0.14	79.87 ± 0.12	80.0 ± 0.52	74.74 ± 0.22
CL + GB + GN (optimal)	Seed = 5,678	76.91	80.08	80.69	74.61
Ours (mean)		77.68 ± 0.10	80.21 ± 0.13	79.31 ± 0.30	76.49 ± 0.18
Ours (optimal)	NR = 0.006	77.98	80.62	80.23	76.48
Ours (optimal)	NR = 0.01	78.02	80.57	79.95	76.70
Ours (optimal)	NR = 0.05	77.92	80.61	80.69	74.61

It is worth mentioning here that we also conducted multiple executions of the experiment by using the same value of NR. Based on our experiment work, we observed

that running the experiment with the same NR did not yield significant changes in performance, as most of the results are repetitive. Thereby, our primary focus was to adjust the values of the NR parameter by using the previously mentioned strategy.

The impact of variation of the NR parameter on model performance can be inferred from Figure 8, which shows the relationship between accuracy and different values of the NR parameter. The figure highlights the interval between 0.001 and 0.05 as, based on our experiments, it induced the highest performance. In fact, within the considered range, it can be observed that there is a presence of a multiple local optimum values. Adjusting the NR parameter is crucial for optimising performance, with a preference for smaller ratios to prevent the distortion of raw signals when large noise is introduced.

From Table 6, we can observe that our technique provides the best performance in classifying PPG samples. In particular, both the application of CL and the addition of GN to the training set produced improved accuracy and F1-score compared to the application of the GB algorithm only. Here, it is noteworthy to mention that generating other values of NR may provide different results. In addition, the combined use of CL and GN also improved the performance of the GB algorithm in terms of sensitivity and specificity.

Furthermore, we can claim that our work improves the performance of classifying PPG data in comparison with the previous work achieved by Lombardi et al. (2022b) with respect to all the four evaluation measures: accuracy, F1-score, specificity, and sensitivity. It is clear that GB with GN allows to achieve competitive performance with respect to the Specificity metric. If we want to improve accuracy and sensitivity measures, we can apply our technique with setting NR to 0.01. Otherwise, if we set NR to 0.05 we can deliver exceptional performance in terms of F1-score. Additionally, we evaluated performance of applying well-known deep learning models: ResNet (He et al., 2016), ZFNet (Zeiler and Fergus, 2014), and ShuffleNet (Zhang et al., 2018) to the used dataset (feature set).

We want to highlight that we compared our work with only one previous study because we exclusively utilise specific PPG data that has been selected from the public database. Whereas, other related works used different datasets for sepsis detection. Thereby, we refrain from conducting further comparisons with other related works as it would not result in a sufficiently fair comparison.

5.4 Threats to validity

In the realm of scientific inquiry, we must acknowledge that the application of the technique presented herein to alternative PPG datasets exhibiting distinct distributions does not offer unequivocal conclusions regarding its performance. Regrettably, the exploration of this limitation was hindered by the unavailability of other pertinent public datasets for comprehensive analysis. Therefore, delving deeper into this subject presents an enticing opportunity for future investigations.

Furthermore, our experimental findings have uncovered a significant sensitivity of the results to the initial values employed for setting hyperparameters in the chosen models. To ensure fair and unbiased comparisons, we opted to utilise the default settings across all models during our study. As a result, the use of other settings when adjusting hyperparameters can change the results of the experiment. Also, generating other GN values can give different results since we use a random generator.

6 Conclusions and future work

This work presents a technique to improve the performance of sepsis detection in PPG data. The performance of the technique presented is improved by employing CL with gradient boosted decision trees. Our technique improves the performance of developing the automated system that detects sepsis in the ICU. The presented technique can examine PPG signals to detect patterns signaling the onset of sepsis, prompting swift action by medical teams to care for ICU patients. In addition, our work offers the possibility of using noise generation with PPG data. Our results show that using GN improves the performance of PPG classification data. Moreover, our work results in better performance compared to the previous work presented by Lombardi et al. (2022b) without using deep learning.

This research opens up several avenues for extension and exploration. One intriguing direction involves demonstrating the performance of the proposed technique on various datasets beyond the one studied here. This will provide valuable insights into its versatility and applicability in diverse scenarios. In addition, there is significant potential in exploring the effectiveness of the suggested approach across a wider range of machine learning models in this scientific study. By undertaking such assessments, we can acquire a more profound comprehension of its resilience and its potential for extensive implementation.

Furthermore, an interesting aspect worth exploring is the evaluation of multiple random configurations during the initialisation of hyperparameters for the selected models. In the same context, evaluating performance of using optimisation methods for tuning NR and seed parameters can reveal overall optimal performance with our presented technique. By doing so, we seek to offer valuable insights to practitioners, helping them achieve the most optimal results possible in their experiments. Future work could also be extended by studying performance of applying more methods of feature extraction. Through exploring these diverse paths, we can deepen our comprehension of the technique's potential and, in turn, bolster its feasibility and efficacy when applied to real-world scenarios.

References

- Aşuroğlu, T. and Oğul, H. (2021) 'A deep learning approach for sepsis monitoring via severity score estimation', *Computer Methods and Programs in Biomedicine*, January, Vol. 198, p.105816.
- Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., Liu, H., Schultz, T. and Gamboa, H. (2020) 'TSFEL: time series feature extraction library', *SoftwareX*, January–June, Vol. 11, p.100456.
- Bisong, E. et al. (2019) *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Springer, Ottawa, ON, Canada.
- Cabanero-Gomez, L., Hervas, R., Gonzalez, I. and Rodriguez-Benitez, L. (2021) 'EEGLIB: a Python module for EEG feature extraction', *SoftwareX*, July, Vol. 15, p.100745.
- Calvert, J.S., Price, D.A., Chettipally, U.K., Barton, C.W., Feldman, M.D., Hoffman, J.L., Jay, M. and Das, R. (2016) 'A computational approach to early sepsis detection', *Computers in Biology and Medicine*, 1 July, Vol. 74, pp.69–73.
- Charlton, M., Sims, M., Coats, T. and Thompson, J.P. (2017) 'The microcirculation and its measurement in sepsis', *Journal of the Intensive Care Society*, Vol. 18, No. 3, pp.221–227.

- De Backer, D. (2023) 'Novelties in the evaluation of microcirculation in septic shock', *Journal of Intensive Medicine*, Vol. 3, No. 2, pp.124–130.
- Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., Shimabukuro, D., Chettipally, U., Feldman, M.D., Barton, C. et al. (2016) 'Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach', *JMIR Medical Informatics*, Vol. 4, No. 3, p.e5909.
- Devlin, B., Fienberg, S.E., Resnick, D.P. and Roeder, K. (1997) *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve*, 1 June, Vol. 14, No. 6, p.251, Springer Science & Business Media, Cambridge, UK.
- Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C.M., French, C., Machado, F.R., McIntyre, L., Ostermann, M., Prescott, H.C. et al. (2021) 'Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021', *Intensive Care Medicine*, Vol. 47, No. 11, pp.1181–1247.
- Fleuren, L.M., Klausch, T.L., Zwager, C.L., Schoonmade, L.J., Guo, T., Roggeveen, L.F., Swart, E.L., Girbes, A.R., Thorat, P., Ercole, A. et al. (2020) 'Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy', *Intensive Care Medicine*, March, Vol. 46, No. 3, pp.383–400.
- Friedman, J.H. (2001) 'Greedy function approximation: a gradient boosting machine', *Annals of Statistics*, Vol. 29, No. 5, pp.1189–1232.
- Gambera, M. (2023) *How to Extract Features from Signals* [online] <https://matteogambera.medium.com/how-to-extract-features-from-signals-15e7db225c15> (accessed 13 August 2024).
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.
- Heo, S., Kwon, S. and Lee, J. (2021) 'Stress detection with single PPG sensor by orchestrating multiple denoising and peak-detecting methods', *IEEE Access*, Vol. 9, pp.47777–47785.
- Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A. and Mark, R.G. (2016) 'Mimic-III, a freely accessible critical care database', *Scientific Data*, Vol. 3, No. 1, pp.1–9.
- Kam, H.J. and Kim, H.Y. (2017) 'Learning representations for the early detection of sepsis with deep neural networks', *Computers in Biology and Medicine*, 1 October, Vol. 89, pp.248–255.
- Lombardi, S., Partanen, P. and Bocchi, L. (2022a) 'Detecting sepsis from photoplethysmography: strategies for dataset preparation', in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, pp.2286–2289.
- Lombardi, S., Partanen, P., Francia, P., Calamai, I., Deodati, R., Luchini, M., Spina, R. and Bocchi, L. (2022b) 'Classifying sepsis from photoplethysmography', *Health Information Science and Systems*, Vol. 10, No. 1, p.30.
- Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C. and Chen, S.A. (2021) 'Neurokit2: a Python toolbox for neurophysiological signal processing', *Behavior Research Methods*, pp.1–8.
- Metz, C.E. (2008) 'ROC analysis in medical imaging: a tutorial review of the literature', *Radiological Physics and Technology*, January, Vol. 1, No. 1, pp.2–12.
- Mollura, M., Mantoan, G., Romano, S., Lehman, L-W., Mark, R.G. and Barbieri, R. (2020) 'The role of waveform monitoring in sepsis identification within the first hour of intensive care unit stay', in *2020 11th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*, IEEE, pp.1–2.
- Moody, B., Craig, M., Johnson, A., Kyaw, T., Moody, G., Saeed, M. and Villarroel, M. (2020) *The MIMIC-III Waveform Database Matched Subset* [online] <http://physionet.org> (accessed 13 August 2024).
- Northcutt, C., Jiang, L. and Chuang, I. (2021) 'Confident learning: estimating uncertainty in dataset labels', *Journal of Artificial Intelligence Research*, Vol. 70, pp.1373–1411.

- Rangel-Frausto, M.S., Pittet, D., Costigan, M., Hwang, T., Davis, C.S. and Wenzel, R.P. (1995) 'The natural history of the systemic inflammatory response syndrome (SIRS): a prospective study', *Jama*, Vol. 273, No. 2, pp.117–123.
- Rudd, K.E., Johnson, S.C., Agesa, K.M., Shackelford, K.A., Tsoi, D., Kievlan, D.R., Colombara, D.V., Ikuta, K.S., Kissoon, N., Finfer, S. et al. (2020) 'Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study', *The Lancet*, Vol. 395, No. 10219, pp.200–211.
- Sakr, Y., Jaschinski, U., Wittebole, X., Szakmany, T., Lipman, J., Namendys-Silva, S.A., Martin-Loeches, I., Leone, M., Lupu, M-N., Vincent, J-L. et al. (2018) 'Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit', in *Open Forum Infectious Diseases*, Vol. 5, p.ofy313, Oxford University Press, USA.
- Scherpf, M., Gräßer, F., Malberg, H. and Zaunseder, S. (2019) 'Predicting sepsis with a recurrent neural network using the mimic III database', *Computers in Biology and Medicine*, October, Vol. 113, p.103395.
- Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J-D., Coopersmith, C.M. et al. (2016) 'The third international consensus definitions for sepsis and septic shock (sepsis-3)', *Jama*, Vol. 315, No. 8, pp.801–810.
- Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P. and Thijs, L.G. (1996) 'The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure', *On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine, Intensive Care Med.*, July, Vol. 22, No. 7, pp.707-710, DOI: 10.1007/BF01709751, PMID: 8844239.
- Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B. and Yang, M-H. (2023) 'Diffusion models: a comprehensive survey of methods and applications', *ACM Computing Surveys*, Vol. 56, No. 4, pp.1–39.
- Zeiler, M.D. and Fergus, R. (2014) 'Visualizing and understanding convolutional networks', in *Computer Vision—ECCV 2014: 13th European Conference, Proceedings, Part I*, Springer, Zurich, Switzerland, 6–12 September, Vol. 13, pp.818–833.
- Zhang, X., Zhou, X., Lin, M. and Sun, J. (2018) 'Shufflenet: an extremely efficient convolutional neural network for mobile devices', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.6848–6856.