Early prediction of sepsis using deep learning algorithm

MSc Research Project

MSc in Data Analytics – Section A

Terrance Thomas

Student ID: 18184928

School of Computing

National College of Ireland

Supervisor: Noel Cosgrave

Contents

Early prediction of sepsis using deep learning algorithm

Terrance Thomas
18184928

Abstract

Sepsis is a life-threatening condition caused by the reaction to an infection from your body. The immune system defends you from many diseases and pathogens, but in response to an infection, it can also go into overdrive. The most common and expensive cause of mortality in critically ill patients is sepsis. Sepsis has about 5.4 million deaths annually. Early detection and treatment of sepsis are essential for patient survival as each hour of delay leads to an average decrease in survival of 7.6%. For effective treatment, the early diagnosis of sepsis is significantly necessary. Tools for encouraging informed decision making will help classify people at the greatest risk for potential sepsis. Machine learning (ML) models may help recognize possible clinical variables and provide greater accuracy than current low-performance conventional models. Several methods have been developed to monitor patients electronically for severe sepsis, but few provide predictive capabilities to allow early intervention. New analytical and ML techniques able to control the wide range of variables that can be used for early detection of sepsis.

Keywords**:** *Sepsis, Machine Learning, Deep Learning, Intensive Care Unit, Predictive Modelling, Medical Computing, Patient Diagnosis, Patient Treatment, ICU patients, Sepsis Diagnosis, Feature Extraction, Sepsis Prediction.*

1 Introduction

1.1 Background

Sepsis is described as severe as “life-threatening organ dysfunction caused by a dysregulated host response to infection”. Sepsis is the third-highest mortality disease in intensive care units (ICU) and has expensive treatment costs. Sepsis occurs when an infection that you already have in your skin, lungs, urinary tract, or elsewhere, induces a chain reaction in your body. Sepsis can rapidly lead to tissue damage, organ failure, and death without early care. In 2013, over 23 billion US dollars have spent in hospitals ICU. (Li, et al., 2019)

Each year around 5.8 million people worldwide die from trauma, 40 percent of deaths happened during hospitalization, 22 percent of which were caused by sepsis. Early diagnosis of sepsis for patients with trauma will allow doctors to intervene and assess care and patient results in advance. Sepsis has become a global public health concern, due to high incidence, mortality and complex pathogenesis. Delayed diagnosis per hour is associated with a rise in mortality of about 4-8 percent. Early sepsis diagnosis will effectively decrease the occurrence of post-traumatic complications and mortality (Fu, et al., 2019).

ICU outcome prediction models are focused on broad population analysis and often provide statistically reliable outcomes for an average patient, but are often costly, time-consuming and vulnerable to bias in selection. Also, these measures are typically lacking the accuracy needed for use at the individual level, as they pose substantial errors in patient data that are far from the average (Garcia-Gallo, et al., 2019).

ICU patients are the most closely monitored patients in the entire hospital; for this reason, ICU is an area rich with data, even to the point of exhaustion. The large volume of data collected from a single patient in an intensive care unit makes it humanly impossible to coordinate and analyse it at the necessary time. Hence Machine learning (ML) techniques are used to enhance efficiency in predicting sepsis (Garcia-Gallo, et al., 2019).

ML is the scientific analysis of algorithms and mathematical models employed by computer systems to perform a particular task without using instructions, relying on trends and inferences. ML is closely connected to computational statistics that focus on computer-based predictions. ML algorithms create a mathematical model based on sample data, known as training data, such that predictions or decisions can be made without specific programming to perform the function. ML focuses on designing computer programs that can access data and learn how to use it for themselves. It is an artificial intelligence (AI) sub-set that gives systems the ability to learn and develop automatically from experience without specific programming. Data mining (DM) is a field of study within ML and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, ML is also referred to as predictive analytics (PA).

The primary objective is to allow computers to automatically learn without human interference or assistance and to adapt actions accordingly. The learning process starts with observations or data, such as examples, direct experience, or feedback, to search for trends in the data and make informed decisions in the future based on the examples we have. ML helps to analyse large quantities of data. While it usually provides quicker, more reliable results to identify lucrative opportunities or dangerous threats, it may also take additional time and resources to properly train it. Combining ML with AI and cognitive technology will make the processing of large quantities of information much more efficient.

1.2 Problem Statement

There is a large amount of medical records of ICU patients available on sepsis and there is no proper treatment for it. Hence, early prediction of sepsis is very crucial. Delay in prediction leads to increase in mortality and high treatment cost.

1.3 Research Questions

The following questions can be answered by the research project

- How can different ML techniques be used to predict sepsis in an early stage?
- To what extend the patient’s medical record can be used to enhance the model?

1.4 Motivation

Premature babies are at high risk of developing sepsis-related infections-a life-threatening and multi-organ complication induced by an immunological reaction to the infection. Premature babies are at high risk of developing sepsis-related infections-a life-threatening and multi-organ complication induced by an immunological reaction to the infection. Sepsis is a major public health concern and contributes significantly to neonatal morbidity and mortality (Joshi, et al., 2020).

Sepsis survivors can suffer from long-term physical and psychological problems requiring more social and health care support (Li, et al., 2019). Despite comprehensive studies in this area, the management and diagnosis of sepsis remain challenging (Amiri, et al., n.d.). Current clinical sepsis detection involves tradition-based approaches that have a lengthy processing period or patient response tracking that adds uncertainty to the therapy. Such deaths consist primarily of elderly, neonates and critical-care patients with impaired immune system. A 2017 world health organization study indicated that infectious diseases caused 50 percent of all deaths in children under the age of 5, more than 0.4 million of whom died in 2015 from neonatal sepsis (Shah, et al., 2019).

Since, the people who are more prone to sepsis are infants and elderly who done have a strong immune system it is even more crucial to predict sepsis so that they can be treated at early stages and not at a stage where treatment is not possible.

2 Related Work

2.1 Literature Review

A bidirectional long short-term memory (LSTM) network which is based on neural networks (NN) to simultaneously combines analyses of various organ systems and intuitively represent the patient's condition in a timely way was developed. There are 3 datasets that are used for the research. 48 features were used from the datasets for the prediction. Four key components of the model are pretraining, self-attention, Bidirectional LSTM, and attention. The model learns the unique features from human organ systems in time series electronic health records (EHR) but also uses the temporal correlation among the organ systems.

All the extracted features were converted into a matrix with multiple numbers of rows. Missing values were replaced with mean values. Multitask deep neural network (MT- DNN) was used for encoding the data. Self-attention vectors capture the meaning of time series data of human organs. Bidirectional LSTM network consists of both forward and backward sub LSTM. Attention mechanism to learn important features selectively and to create direct short connections between the target and the source.

The proposed model achieves 94.72 percentage of F1 score on one of the datasets. However, the scores decrease over the next datasets. A better method can be proposed to improve the accuracy which is better and performs same on all the datasets (Li, et al., 2019).

The research focuses on early sepsis prediction by using a ML model which is an improved cascade deep forest model. The model worked on the features extracted from the patient’ s Electronic medical record (EMR) of the first 24 hours of the ICU admission. The dataset had 3125 patient records of which 1187 patients had sepsis while the remaining did not have sepsis. The records of patients were of adults i.e. all above the age of 18 and any ICU admission which was less than 8 hours was not taken into consideration.

Feature extraction was used on the dataset. 3 set of data was available viz demographic, vital and laboratory data. Initially, 105 features were extracted however only 85 features were used. The model achieved Area under the receiver operating characteristics (AUROC) of 0.80, sensitivity of 0.79 and specificity of 0.64 which was better than all the baseline models used like support vector machine (SVM), random forest (RF), logistic regression (LR), k-nearest neighbours (KNN), gradient boosting(GB) and extreme gradient boosting (XGBoost). However, the prediction is not real-time, and the dataset used was small. Also, additional data like images could be used in combination with text data to build a more precise model (Fu, et al., 2019).

The model developed is an approach that focuses on the evaluation of the impact and significance of three separate patient similarity measures on the estimation of one-year mortality in patients with the same diagnosis as the traditional predictive models are based on large-scale population analysis and can provide statistically reliable results for an average patient, but are often costly, time-consuming and vulnerable to selection bias; however, these measures typically lack the precision needed for use at the individual level, as they pose substantial errors in the patient data that are far from the average.

Four criteria were used to identify sepsis patients. First, patients with ICD-9 diagnosis codes: 995.92 for severe sepsis and 785.52 for septic shock. Second, a validated protocol that uses administrative data to identify sepsis patients. Third, patients were sorted either by codes for septicemia, bacteremia, disseminated fungal infection, disseminated candida infection or disseminated fungal endocarditis in addition to an organ dysfunction code. Fourth, sepsis admission based on the most recent definition of sepsis 3.

Of the 16,219 data, only 16,080 considered. Five specific measures of patient similarity to pick the best outcome with respect to the one-year estimate of mortality model. The best performing model which used LR algorithm presented an AUROC of 0.83 and the model that used all available data for training presented an AUROC of 0.81. To get better results implementation of novel ML methods on graph-structured data such as a convolutional neural network (CNN) and assessment of various similarity steps had been suggested (Garcia-Gallo, et al., 2019).

The study emphasis on premature infants was planned to identify the prognostic potential of multiple heart rate variability (HRV), respiration, and (infant) motion characteristics for the predictive monitoring of late-onset sepsis (LOS). Sepsis in neonates can be characterized into 2 early-onset sepsis (EOS) and LOS. They are characterized into 2 categories by the timing of the symptoms, severity of infecting organs and associated pathogens. EOS shows symptoms within 72 hrs of birth while LOS shows symptoms after 72hrs of birth. Prevalence chances are higher in LOS than EOS almost by 15% of infants who are weighed under 1.5 kg. Diagnosing LOS is more challenging as the clinical signs are nonspecific and not clearly visible.

R peak was recorded from the ECG electrodes (ECG and CI waveform). Signals used for the analysis included HRV, CI-based respirational signal, and ECG waveform. 8 features of HRV was considered, 11 features of CI-based respiration were taken and 3 features from ECG has been used. Because baseline values of the characteristics that vary from infant to child, the time series of all characteristics have been standardized due to factors such as differences in physiological maturity.

Naive bayes (NB) classifier was used on the data set. NB was considered as works well on small data set and accuracy is high. The true positive ratio (TPR) to false positive ratio (FPR) was 10%. The root mean square of the successive differences (RMSSD) was 66% and the average acceleration response was 61%. However, the model does not involve the study of infants with a negative blood culture as well as model validation for predictive and the dataset size was small. Also, the model predicts the relative risk of sepsis and not the actual risk of sepsis in the hours leading to sepsis (Joshi, et al., 2020).

A methodology was developed to allow automatic sepsis prediction 6 hours before its clinical presentation. Four vital signs of heart rate (HR), spontaneous bacterial peritonitis (SBP), temperature and respiratory rate are scored along with laboratory tests for platelets, white blood cells (WBC), glucose and creatinine. The weighted scores obtained from the screening methods are also used to categorize patients into 4 classes of separate ICU sepsis probabilities.

Data set consists of 40,336 ICU patients with 40 different time variables which include Vital Signs, Laboratory (VSL) values and Demographics. Screening tools are used for early detection of sepsis in both medical and surgical patients are commonly used as decision support systems. Two screening tools of prehospital early sepsis detection (PRESEP) and sequential organ failure assessment (SOFA), as well as the systemic inflammatory response syndrome (SIRS) criteria, were considered. For the missing values, the average of each variable over the time of stay in ICU is considered as the defining feature of that variable per patient.

To overcome the biased dataset, under-sampling had to be performed by using the mean features of the 8 variables considered for each patient to have been scored using PRESEP, SIRS and SOFA screening tools. The new data set under-sampled is grouped into 4 classes, considering the average scores of the three screening methods and the additional ranking. The model used KNN classifier gave a prediction of 96% but gave less accuracy when datasets of other hospitals were used (Li, et al., 2019).

A system was proposed that focuses on identifying optimal HRV features for early identification of sepsis in intensive Pediatric care. Demographic details (weight and age) of suspected patients, date/time of admission to ICU, and the date/time of the results of the infection culture were reported. They were also followed up during the patient's presentation in these units, and their ECG signal was continuously collected.

HRV signals were measured for all selected patients, removing high-frequency peaks and any other unnecessary noise in the signal using pre-processing techniques. The clean HRV signal extracted 28 linear and nonlinear characteristics. The relative entropy selection function algorithm was used to select the best features from among all the extracted features.

Four models viz DT, SVM, KNN, and linear discriminant analysis (LDA) classifiers were applied. DT had the best accuracy of 86.36% among all the models applied. Results are consistent with the physiological findings in this field showing the sympathetic nervous system and critical role in the initial sepsis cascade in the body and one of the key effects of this activation is to increase the heartbeat frequency. By using ML techniques, sepsis can be detected up to 24 hours before clinical diagnosis. The results may be further checked using the greater size of a dataset. New features that may be more reflective of the HRV pattern during sepsis activation in the body are suggested to be examined (Amiri, et al., n.d.).

Septiflo was a device created which focuses on t providing a concentration-dependent qualitative estimation of the sepsis-associated burden of bacterial infection in the blood below 10 minutes. Pre-processing of the images includes a primary test is conducted to determine the image's validity, i.e. whether the image is fully illuminated and whether the cartridge is correctly inserted after the preparation stage. If the cartridge passes the initial test, then the centroid image will be detected where the detected coordinates (X, Y) will be taken for further comparison. danielsson's theory of distance mapping is used for identification of the central point, the region where the actual reaction occurs. As part of device settings, the max & min potential predicted circle radii are fed. It calculates the sum of the strength of all the pixels and the number of pixels from the thresholded image. The measured mean strength of pixels (CMPI) is the average of those two parameters.

The average reproducibility variation in a single instrument is 9%. The reproducibility was estimated to be less than 8.2% between the two instruments. The correlation obtained between the endotoxin concentration and measured CMPI with an R-square value of 0.97 is very important. The colorimetric reader developed represents an accurate, precise, portable and user-friendly reader system for fast and rapid diagnosis of sepsis (Shah, et al., 2019).

The research compares different ML tool for early prediction of sepsis and provides novel imputation and selection of features based on a signal processing system. The goal is to continuously update the projected probability that the experience will result in sepsis, using all available knowledge up to that point. There were 40, 336 subjects in the dataset. The data included demographics, vital signs, laboratory values, sepsis onset time, and sepsis mark for each subject. Continuous variables were logarithmically transformed to reduce the influence of z score.

For ML algorithms, sliding window features are used to access local and global ICU data descriptors for a range of window sizes. 18 features were selected including moment statistics on the distribution of waveforms (mean, variance, skewness, kurtosis) and quantile information. For NN methods, all 40 variables are utilized. The input is the data stored, with imputation and interpolation involved.

The research concluded that NN, RF, sparse quantile regression, NB, and the neighbourhood methods give superior precision, sensitivity, and specificity efficiency. In comparison with the other algorithms, RF provides deceptively high performance on average, but the sensitivity is very weak. In comparison, sparse quantile regression outperforms other sepsis detection algorithms and is resistant to overfit (Hsu & Holtz, 2019).

A model is developed that predicts the onset of sepsis early by observing the missingness of physiological variables and comparing them with general data patterns. The dataset used was a combination of 3 datasets. It has been shown that patients with sepsis and non-sepsis have major variations in ICU duration of stay. Nearly all patients who had not been diagnosed with Sepsis spent less than 60 hours in the ICU (mean stay of about 37 hours). Patients with Sepsis, on the other hand, spent more time at ICU (mean stay of around 60 hours).

Inspired by the importance of informative missingness (IM) posed by the results of these earlier works, data was analysed in three stages to search for variables that show IM. First, patients were divided into groups of Sepsis and non-Sepsis. Then estimated the average observation rates for each variable (with the exception of those with 100% observation) and compared them for the 2 groups. Second, patterns were observed in the variables defined as important (IM variables) to patients with Sepsis. Hourly probability of observation of these variables was plotted for both groups. There were persistent differences in probabilities for observation among groups. Hourly probability of observation of these variables was plotted for both groups. There were persistent differences in probabilities for observation among groups.

Researchers were most interested in a time-optimal 6 hours' time period. There is a marked increase in the likelihood that both variables will be observed a little after time optimal. The data was based on the XGBoost algorithm. The base model is a naive model simply used a windowed input was called XGBoost-W. There was a modification made to the model. The "best model" described in the paper which is XGBoost-W and a larger masking vector was generated using all the temporal variables, further increasing the Utility score (All IM). The best model gave an AUROC score of 0.8406 and mean Utility of 0.404 on 5 fold stratified cross-validated data (Singh, et al., 2019).

The study focuses on improving efficiency and proposing a new approach with better results using low dimensional algorithm. The dataset record has 40 variables, the first 34 of which are clinical, and the rest are demographic. The parameters are available on an hourly basis in each record. Low dimensional algorithm is used which consists of six parameters, redefined weightage, and criteria.

After the data was read, the data were reviewed for the Parameters and Assign weightage. This helped filter out the data according to the appropriate weighting parameters specified by the parameters chosen, and the threshold criteria applied to the distinction between patients with sepsis and non-sepsis. In any sepsis detection procedure, the parameters are given some weight to achieve quantitative test sepsis. The overall weighting was tested based on these parameters and weighing. If the total weight is > 2 then the patient has sepsis, then the patient has no sepsis. Now the latest procedure repeats for the next data hour. The 3 datasets were transferred, and each dataset's results are 0.968, 0.978 and 0.984 accuracy. Even though the accuracy was high, the model failed to detect sepsis in early stages due to the missing data which limited the detection (Deogire, 2019).

Using bagged DT a model is developed to predict sepsis. Each patient's multi-function record consists of many samples of features usually gathered every hour. Features were pre-processed in several steps including the elimination of outer, combination of associated variables, imputation of missing meaning and standardization. The pre-processed records were then divided into data set for training and validation. Using the training data set a three-bagger classifier was trained and the appropriate features and hyperparameters were selected in an iterative process until the best score was reached. The classifier was validated by validation data collection, and the model was submitted to test a hidden dataset subset value. The classifier with the best utility score was selected on the full test dataset for evaluation of the final utility score.

Using the binary outputs, a group of DTs was developed: sepsis or no sepsis. The trained model accepts the basic features as input over time for each sample. The unit consists of 100 DT. Due to the highly unbalanced nature of the data, a cost ratio of 1 to 37 was developed for no-sepsis versus sepsis. The maximum number of splits with a minimum size of three leaves is set at 100. AUROC was 0.764, 0.768 and 0.741 on the 3 datasets. The low score may have resulted might be due to the algorithm's training using the features 'single time samples and did not acknowledge the association with previous samples (Firoozabadi & Babaeizadeh, 2019).

A model using time stacked groups and ML algorithms to predict early sepsis diagnosis. The main goal was to optimize a particular utility feature with a reward for predicting sepsis 12 hours before and 3 hours after the initiation of the sepsis and particular penalties for false negative and false positives. The dataset includes 1,552,210 data points for 40,336 patients.

For compute quantiles, quantile ranges, and discrepancies and quotients to the actual value, rolling periods of 6, 12, 24, and 48 hours were introduced. 'last observation carried forward process' to copy the last available laboratory result and critical parameter to the actual date if there is no more recent data. This method reflects the medical viewpoint of decision making and typically calculates the laboratory findings from blood samples with varying frequencies. Missingness was made clear by the use of binary variables to show that the values were forwarded. Numerical variables representing the actuality of the given value (0 for new measurements, 6 for measurements 6 hours ago) so that machine learning models could learn the significance of outdated variables.

A non-specific XGBoost learner for predicting sepsis 6 hours before the onset of sepsis. Utility score of 0.394 for the XG-Boost model which is not unique. In addition, 86.2 percent of non-sepsis data were correctly reported, resulting in AUROC 0.823. Time-specific meta-learners showed potential for a deeper understanding of the driving forces that characterize the pre-septic condition, depending on the hour after ICU entry. Commonly used clinical scores lose their ability to screen for sepsis after the first day of ICU admission, which shows the need for better scores for regular screening, particularly in sepsis acquired by ICU (Vollmer, et al., 2019).

A sepsis prediction model is developed based on the clinical time series results. The proposed method uses a LSTM network of deep professional architecture and residual connections. Missing values are among the most difficult problems considering data from clinical time series. Advanced predictive possibilities for missing values include regression techniques such as LSTM. The standardization function is extended to the fixed range of values by replacing missing values with numerical representations from outside the standardized range.

The sepsis occurs among all patients at 1.80 percent of all time points. The septic diagnosed number of patients is 7.27 percent. A small number of sepsis class members cause difficulty in most of the traditional prediction methods, leading to non-sepsis class prediction in all situations. This challenge can be solved by oversampling of the rarer (septic) class or using higher weights for rarer level samples. The network proposed consists of 7 sections, each consisting of an LSTM layer followed by 3 fully connected layers.

The network performance is a predictive score of sepsis at each point in time. The final best model uses 7 lines, with substitution of normalization/failure value. This results in 0.350 utility and 0.131 dice and 0.372 for dataset A. The proposed approach results in a standardized utility score of 0.281 for the complete test range. The proposed method addresses all the issues with early detection (Vicar, et al., 2019).

The model works on facial representation Learning for Septic Shock Early Prediction. A general framework that represents a coherent and systematic way, the derived temporal relationships and local patterns through facial representations that have changing emotional expressions based on patient health conditions. This type of feature representation not only enhances the visualization potential of EHRs but also further benefits the downstream task of predicting septic shock early.

For the first time, both static and dynamic information is processed through the Image Generator module which produces a sequence of facial images. The images produced are then fed into the Prediction module. 2D CNNs are used to extract high-level facial representation characteristics from the produced images and then fed the extracted facial representations into LSTM based models to predict the visit level.

At the heart of the proposed system is the Image Generator, where the EHRs are ingested as their input and a sequence of facial representations is generated with different emotions. A mapping scheme in which the four sepsis stages into four corresponding distinct emotions: "neutral" to "infection" because the sample population is patients with suspected infection, "sad" face to "inflammation," "angry" face to "organ failure," "fear" face to "sepsis shock". A fifth, a "happy" emotion for "recovery" or not at any of the four sepsis stages.

Both of two proposed models surpass all eight baselines for all tests, with one exception being that of img.2D CNN.LSTM + sta. Specifically, comparing img.2D CNN.LSTM + sta to dyn. LSTM + sta, and found improvement in every metric. This finding suggests that the facial representations produced can indeed capture important information for early prediction of septic shock. This result indicates that while facial representation learning currently distills the most outstanding information for the early detection of septic shock from the original EHRs, the complex information can also contain additional information (e.g. subtle changes in vital signs) that is still predictive but not captured by the facial representations. And mixing the two is advantageous. The initial EHRs are highly incomplete, plagued by missing values and irregularity in time. The system generates an image, it completes the missing value to some degree by training the components in image generator to predict sepsis stages using relevant information (Lin, et al., 2019).

The research is done of sepsis in infants and how accurate it is. The data set of infants included generic information about the patients and lab results. In the data collection, gender, form of birth, birth weight, birth week, actual birth week, mismatch of the blood group, Celestone application, Rhogam application, antibiotic use, RDS, surfactant fields were used.

ML algorithms of KNN classification and the NB classification was used. KNN algorithm allowed 94.53 percent accuracy diagnosis of Sepsis while NB algorithm provided a 93.75 percent accuracy (Tekin, et al., 2019).

A model is made that is based on attention-based interpretable approach to diagnose sepsis. The data collection contains vital signs, laboratory tests, diagnosis codes, codes of treatment, findings and information made by medical personnel and more than 53,000 adult patients. However, a subset of the dataset was only used for the research. In total, 11,791 patients are in the studied cohort. A total of 39 features were selected after feature selection. Filling forward filled the missing values.

The research uses Recurrent Neural Networks (RNN). Gated recurrent unit (GRU) is implemented in the research. Unidirectional RNNs are not used. Instead, bidirectional RNNs used as are. Bidirectional RNNs are a point of interest in the study of the time series. Output has two classes: Patients with septic and non-septic. The efficiency of the model in AUROC reaches over 0.75 precision.

The accuracy can be improved by firstly, consider other evaluating methods for handling the missing data. Second, proper data clustering can potentially result in more interpretability and a lot of reliable performance. Finally, to feed the data into the model has to be researched which will improve the data set's expressiveness that would hopingly produce better performance metrics (Baghaei & Rahimi, 2019).

The focus is on sepsis in patients in the emergency department (ED). The database consists of measurements of vital signs, records of fluid administration and laboratory measurements during the stay of patients. Of 186,000 documents in the database, a cohort of 761 high-risk septic shock experiences. Due to the sparseness and irregularity of measurements of ED vital signs, time series were interpolated linearly between recorded observations at a resolution of one minute.

Mean arterial pressure (MAP) and fluid bolus therapy (FBT) had to be extracted and clustered. Changes in MAP have been studied for analysing FBT responses. The derived MAP time series for each bolus documented within a time range from 15 minutes before to 2 hours after the reported start time for the bolus. Before the vasopressor administration, the relationship between HR and MAP was analysed. The HR and MAP time series is viewed as a trajectory for each patient in the 2-dimensional phase space and divided into consecutive 3 or5-hour time windows of equal size. Time-series clustering of MAP in response to FBT shows that approximately 40 percent of boluses did not increase MAP and that the shift in MAP also seemed to be related to the initial MAP level. The model can be enhanced by adding more variables in clustering such as saturation of blood oxygen and body temperature to help provide practical clinical practice recommendations for handling septic shock hemodynamics (Gu, et al., 2019).

A new model is proposed that diagnosis of Sepsis in early stages in pediatric ICU patients by identifying optimal features from HRV. In the ICU section, patients were monitored continuously for vital signs including ECG signals. All patients who had any outcome of the infection culture test including blood cultures, tissue culture, sputum culture, etc. were monitored. They were identified as suspicious patients and their demographic details (weight and age), the date/time of admission at ICU, and the date/time of the findings of the culture of infection were registered. Also, the ECG signal was continuously collected from the patient.

More than 300 patients were identified as perpetrators who received at least one infection culture test result (blood culture, sputum culture, etc.). The ECG signal was collected 24 hours before clinical sepsis diagnosis. Then, using the ECG r-peaks detection algorithm in matlab software, the extracted the HRV signal for each of these. rlowess filter, an updated variant of the low-frequency system to eliminate any excessive high-frequency peak. This filter attempts to find a smooth curve between the data points without needing any advanced functional relationship specifications among the variables. 28 features were extracted for the research. DT, KNN, LDA, and SVM were applied on the dataset and DT gave the best accuracy of 86.36%. With a larger dataset size, the results may further verify the accuracy of a model. New features that may be more reflective of the HRV pattern can be used during sepsis activation in the body (Amiri, et al., n.d.).

The paper focuses on creating an application that predicts sepsis using CNN on an early stage for preterm new-borns. Real-time vital signs like heart rate (HR), respiratory rate (RR), and saturation of blood oxygen (SpO2) from 32 cots attached 24/7. The dataset contains 146 new-borns. All data files in which the invalid record ratio exceeds 10% is deleted and the invalid record replaced with the mean value. The data had to turn from streams of vital signs into pictures. Sequential data does not work well on CNNs with 2-D filters. Therefore, data chunks were used which are sequences of data cut to a certain length and transformed.

A 14-layer deep CNN was used. The proposed model AUC 0.79 precision 0.76. But the model has certain limitations: Lack of data makes the converted images not large enough, the transition from sequential data to the image is simplistic and only suspected cases are taken into consideration. LSTM can be used along with CNN for better results (Hu, et al., 2019).

The model developed was a non-invasive early detection treatment for neonatal sepsis using HRV tracking and ML algorithms. The goal is to predict the early risk of sepsis within the first 48 hours of a new-born's life. The data set included 79 new-borns. Firstly, the time difference between consecutive R peaks was extracted from the register to beat to beat intervals (IBI). Secondly, a pre-processing stage for correction of the ectopic beat, detrending, and IBI resampling was introduced. Ultimately, it measured time, frequency and non-linear parameters.

Adaptive Boosting (AdaBoost), Bagged Classification Trees (BCT), RF, LR, SVM, NB, Classification Tree (CT), and KNN were applied on the dataset. RF, BCT, and AdaBoost achieved the highest performance (Gómez, et al., 2019).

2.2 Conclusion

From the above papers, it can be seen how the performance of different ML techniques depends on the data set, the cleaning, missing values, and data transformation. More advanced methods of deep learning (DL) can be used to get better results. Some algorithms perform well on small datasets while some require a larger dataset. Some algorithm performs well only on a given dataset; if a similar dataset is given the performance of the algorithm decreases. Also, the pre-processing is a key role for the algorithm selection. Based on the dataset for this research which has a lot of missing values and large data; the LTSM algorithm can be used which performs well for missing data and can handle large datasets as well.

3 Research Methodology

3.1 Introduction

There are many methods of data mining that can be applied to current data science projects. Data are obtained and stored at a dramatic rate across a large variety of fields. A new generation of scientific theories and methods is urgently needed to help people extract valuable information (knowledge) from the rapidly increasing amounts of digital data.

Cross-industry Standard Data Mining Process (CRISP-DM) and Knowledge Discovery in Databases (KDD) are the most common methodology followed. For the research KDD approach is followed to get results by early detection of sepsis.

3.2 Methodology

KDD is concerned with the development of different methods and techniques to make sense of the data. The fundamental problem addressed by the KDD method is that of mapping low-level data (typically too voluminous to easily understand and digest) into other types that may be more lightweight, abstract or useful. The conventional method of translating data into information is based on hand analysis and interpretation. The conventional approach of translating data into information is based on hand analyses and interpretations. The experts then send the report describing the analysis; this report is the basis for future health-care management decision making and planning. In every field, data analyses have become crucial. KDD was developed to overcome the traditional method.[1]

Figure 1 shows the process of KDD flow. For this research, KDD can be used and has 5 important stages. The first stage is called data selection. In this stage, the data is selected which will be used for the research project. However, before the data is selected the purpose and scope are also defined. The second stage is the data pre-processing. Once the data is selected which must be used and if the data does not violate any ethical issues then pre-processing can be done. Pre-processing is a key step in KDD. It involves data cleaning, removing unwanted data, impute missing data. This ensures data is suitable for the algorithm which helps in better performance of the algorithm. The third step is called transformation. Once the unwanted data is filtered out and data is combined now it must be transformed. Data transformation can be transforming the data from one form to another as per the algorithm or the model. It also means transforming data from a complex to a simple format (Fayyad, et al., `996).

[1] https://medium.com/datadriveninvestor/data-science-project-management-methodologies-f6913c6b29eb

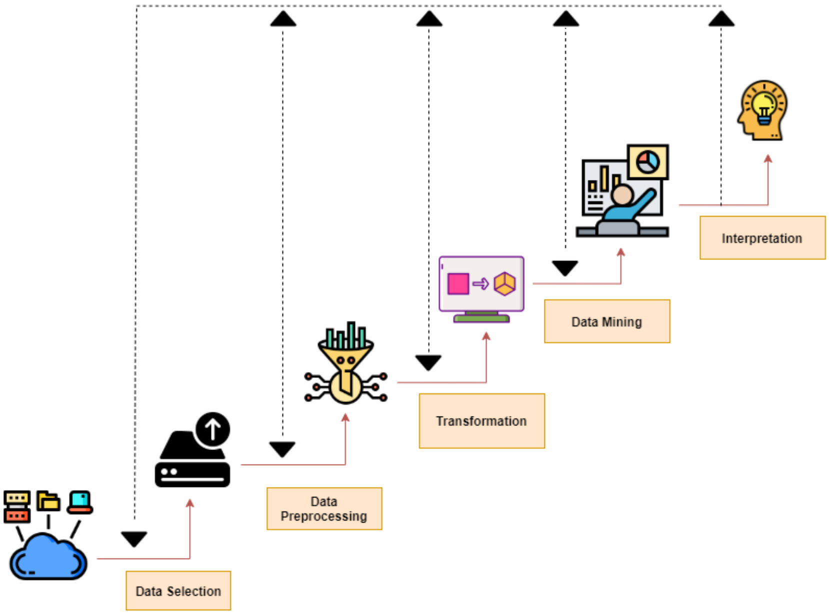


Figure 1: KDD flow

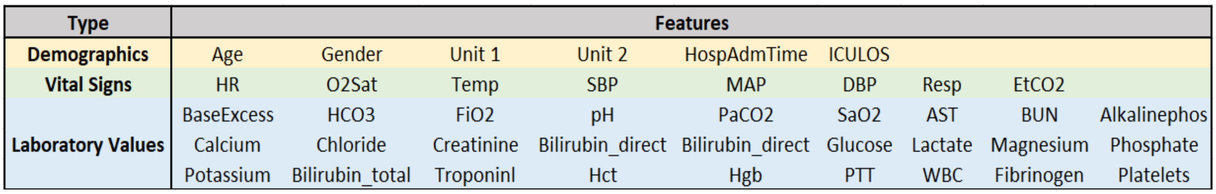
The fourth step is called DM. In this stage, the proposed ML model or algorithm is applied. The dataset is divided into test and train. The training dataset is used to train the model. Once the training is done the model is run on a new data called test which was never used on the model. The fifth stage is called interpretation. This is also a vital stage. The data/information obtained from the model is used to display in a visualization that helps a better understanding of the data. Most important part of KDD is that at any point of time the process can go back to the earlier stages due to the feedback loop which helps if there the data gathered is insufficient or the process skipped or took a wrong decision.

3.3 Proposed Dataset

The dataset used for the research is from a challenge held in 2019 by physionet computing in cardiology. Challenge data are from three geographically distinct U.S. health systems with three separate systems of electronic medical records (EMRs). These data were collected by the required Institutional Review Boards over the past decade with the approval.

The team that conducted the challenge identified and posted the data and labels as public training sets for 40,336 patients from two of the three hospital systems, and sequestered data and labels as secret test sets for 22,761 patients from three hospital systems (Reyna, et al., 2019).

The data comprise 40 clinical variables, including 8 summaries of vital signs, 26 laboratory values, and 6 patient descriptions. Figure 2 shows the details about the data variables in the dataset.



Type	Features								
Demographics	Age	Gender	Unit 1	Unit 2	HospAdmTime	ICULOS			
Vital Signs	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	
Laboratory Values	BaseExcess	HCO3	FiO2	pH	PaCO2	SaO2	AST	BUN	Alkalinephos
	Calcium	Chloride	Creatinine	Bilirubin_direct	Bilirubin_direct	Glucose	Lactate	Magnesium	Phosphate
	Potassium	Bilirubin_total	TroponinI	Hct	Hgb	PTT	WBC	Fibrinogen	Platelets

Figure 2: Dataset variables

3.4 Data Pre-processing

Handling missing values are among the most difficult issues considering data from clinical time series. There are some common solutions to how missing values are treated. The easiest one is a constant value substitution, e.g. zero or average. Another way of using the last known value. Regression techniques can be used to the prediction of missing values which is more advanced. Another solution is the use of the prediction technique which can fix missing values like the NB classifier.

The feature normalisation can be used within the fixed range of values, including replacing missing values with numerical representations from outside the defined range. Values of each function fit within range 1,5 and missing values were replaced by 0. Since LSTM is an advanced method; feature normalization can be used as the Network will identify and learn to accept zeros as non-informative.

Values that correspond to some physiological possibilities were out of the specified range. For such values, it can be replaced by the highest or lowermost limit value depending on which side the value falls on. It is possible to eliminate clinical features that mostly had missing values (Vicar, et al., 2019).

Feature extraction can be also used. It is a reduction of dimensionality by which an initial collection of raw data is reduced to more manageable groups for processing. It is useful when reducing the amount of resources necessary to process without missing essential or significant information. Extraction of features may also reduce the amount of redundant data needed for a given analysis. In addition, data reduction and the machine's efforts in creating variable combinations (features) accelerate the learning speed and generalization steps in the ML process. [2]

3.5 Data Imbalance

The dataset is also highly unbalanced. Sepsis occurs only in about 1.8 % of the entire dataset. Also, the patients diagnosed as septic is 7.27%. It can cause trouble in most standard prediction methods, leading to non-sepsis type prediction in all situations. To address the problem, oversampling of the rarer (septic) class or implementing higher weights for rarer-class samples can be done. The typical approach for NN is the implementation of the weighted loss function, such as weighted cross-entropy or generalized loss of dice (GDL) (Vicar, et al., 2019).

[2] https://deepai.org/machine-learning-glossary-and-terms/feature-extraction

3.6 Data transformation

Data transformation is important especially for the method applied as it helps to change the data format suitable for the algorithm. For the approach in this research, hardly any data transformation is required as the algorithm is an advanced method that can handle most types of data. The proposed loss function does not guarantee the best output score threshold (in terms of the utility value) would be 0.5. A grid quest was used to find the best threshold, where the utility factor for the validation set was maximised.

3.7 Data Mining algorithm

LSTM: LSTM is a type of NN. It is an artificial RNN architecture used in the field of DL. LSTM is different as it has a feedback connection which makes it better than standard feed-forward NN. LSTM can process simple and complex inputs. It is well suited for classification, prediction on time series data. More importantly, it can handle missing data well. The training of data is carried out in a supervised manner and has many variants. LSTM networks have certain internal contextual state cells functioning as long-term or short-term memory cells, in a simple way. This is a very important property because we need the NN prediction to rely on the historical history of inputs, rather than just the very last input.[3]

4 Design Specification

4.1 Proposed Algorithm

The network will consist of 7 sections, each consisting of the LSTM layer followed by 3 fully connected layers. Inspired by the residual neural network (ResNet) and densely connected convolutional networks (DenseNet), it is possible to add residual skip connections, where each block's input is a concatenation of the previous block's output, skip over the previous block's and network data. The rectified linear unit (ReLU) and dropout layer (with a probability of 0.5 drop) are followed by each completely connected block as shown in figure 3. Inputs into the network are all previous time points prior to the time point evaluated. The output is a sepsis / non-sepsis prediction score at an observed time point. The output softmax layer ensures that output values are mapped into range 0-1. [4]

[3] https://medium.com/datathings/the-magic-of-lstm-neural-networks-6775e8b540cd

[4] https://en.wikipedia.org/wiki/Rectifier_(neural_networks)

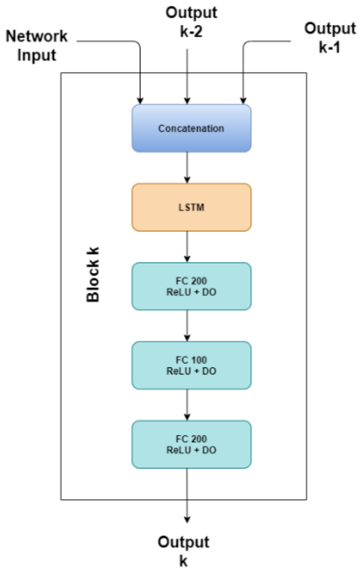


Figure 3: Proposed method flow

4.2 Hold Out Sample

For this research, only the data from the 2 hospitals which were provided for training will be used. The data will be split into 2 parts, the training part, and the testing part. The training part will contain 75% of the data while the testing data will contain the remaining 25% of data.

4.3 Technical System setup

The hardware and software specifications which are recommended for the research implementation is given below in Table 1 and Table 2.

Table 1: Hardware Configuration

Memory	Processor	Processor Speed	Storage Memory
4 GB RAM or more	Intel i5 or equivalent level processor	1.8 GHz or above	10 GB or above

Table 2: Software Configuration

Computing resources	Tools	Libraries
NCI OpenStack	Jupyter Notebook	Keras, Scikit learn
Google Colab TPU	OpenCV	Tensorflow, Theano

Kaggle TPU, GPU Notebook	PyCharm	Pandas, Numpy, Plotly
Amazon S3 storage service	Github (version control)	Pytorch

5 Proposed Evaluation

The objective is to test the model implementation. The metrics which are proposed to measure the results given as follows confusion matrix, sensitivity, specificity, positive predictive value (PPV), F1 score and balanced accuracy (BACC).

The confusion matrix has values of True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP).

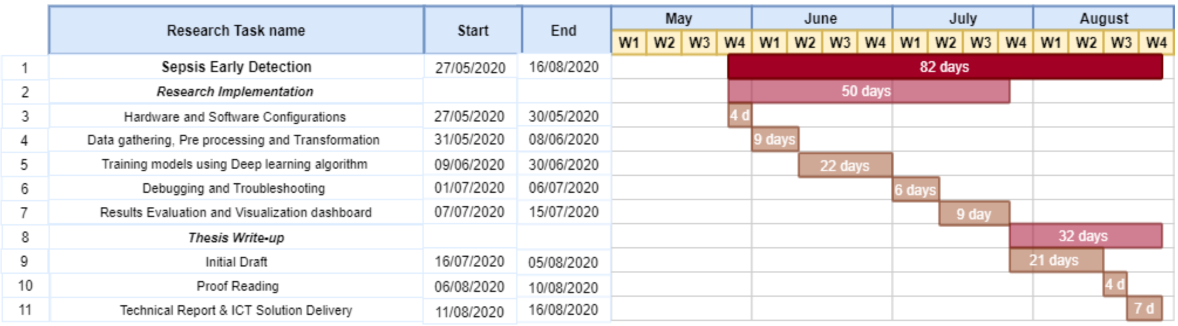
1) Sensitivity (SEN) = TP / (TP + FN)
2) Specificity (SPEC) = TN / (TN + FP)
3) Positive Predictive Value (PPV) = TP / (TP + FP)
4) Balanced Accuracy (BACC) = (SEN / SPEC) / 2
5) F1 score (F1) = 2 / (SEN—1 + PPV—1)

The above suggested evaluation will help determine the accuracy of the proposed model.

6 Proposed Project Plan and Ethics

6.1 Gantt Chart

For any project both as a researcher and in corporate departments, the project undergoes several phases before the solution is provided. The execution of any project requires critical measures such as careful preparation, tracking and progress tracker. A Gantt chart has been plotted for this research to demonstrate the approximate time taken during the third semester for each task as expected.



	Research Task name	Start	End
1	**Sepsis Early Detection**	27/05/2020	16/08/2020
2	***Research Implementation***		
3	Hardware and Software Configurations	27/05/2020	30/05/2020
4	Data gathering, Pre processing and Transformation	31/05/2020	08/06/2020
5	Training models using Deep learning algorithm	09/06/2020	30/06/2020
6	Debugging and Troubleshooting	01/07/2020	06/07/2020
7	Results Evaluation and Visualization dashboard	07/07/2020	15/07/2020
8	***Thesis Write-up***		
9	Initial Draft	16/07/2020	05/08/2020
10	Proof Reading	06/08/2020	10/08/2020
11	Technical Report & ICT Solution Delivery	11/08/2020	16/08/2020

Figure 4: Project plan - Gantt chart

As seen in figure 4, the total project timeline is of 82 days estimate starting from 27 May 2020 till 16 Aug 2020. In this time, two key components expected are "Research

Implementation" and "Thesis Write up" which can be projected to be done in 50 days and 32 days. During device configuration, the plan included troubleshooting, debugging errors if any in implementation level, as well as using Latex for documentation purposes.

6.2 Ethical Implications

The new European general data protection regulation (GDPR) security imposes stringent limits on the collection of personally identifiable data. The GDPR does not only concern European companies, as the regulation extends to all organisations that monitor European citizens or provide services to them. Free exploratory data analysis is only allowed on anonymous data, at the expense of certain legal risks. The GDPR took effect on May 25, 2018, imposing strict limits on the collection of personally identifiable data.

The data used for research is taken from the physionet challenge that the appropriate Institutional Review Boards gathered with the approval over the past decade. It is composed of some basic details and some consumer profiles. Some vital signs and Laboratory values are also collected. The data is openly available to everyone. The dataset does not violate the GDPR rules.

7 Conclusion

Sepsis is a condition that occurs when the body reacts to an infection. Since the body is busy fighting the infection; sepsis worsens the condition which is why is the most common reason for death for ICU patients. Moreover, the treatment of sepsis is expensive and patients even after treatment require help. Even a delay of 1 hour can increase the mortality rate by 7.6% which makes it even more important to detect sepsis as early as possible as infants and the elderly are more prone to it. Using ML algorithms, the data of the patients can be used to predict if there is a possibility that the patient might get sepsis. Using LSTM which is a part of DL to predict sepsis for a large dataset which has about 40 variables and which has a lot of missing value. This algorithm if performs well then can be a key to early prediction of sepsis for patients who have missing data. Also, if the data is available the same algorithm can be used as well.

Bibliography

P. Amiri, A. Derakhshan, B. Gharib, Y. H. Liu and M. Mirzaaghayan, "*Identifying Optimal Features from Heart Rate Variability for Early Detection of Sepsis in Pediatric Intensive Care*," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 1425-1428..

Baghaei, K. T. & Rahimi, S., 2019 . "*Sepsis Prediction: An Attention-Based Interpretable Approach*". New Orleans, LA, USA, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE, pp. 1-6.

Deogire, A., 2019. *A Low Dimensional Algorithm for Detection of Sepsis From Electronic Medical Record Data*. Singapore, IEEE, pp. 1-4.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., `996. *"From Data Mining to Knowledge Discovery in Databases"*. AI Magazine, 17(3).

P. Amiri, A. Derakhshan, B. Gharib, Y. H. Liu and M. Mirzaaghayan, "*Identifying Optimal Features from Heart Rate Variability for Early Detection of Sepsis in Pediatric Intensive Care*," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 1425-1428.

Firoozabadi, R. & Babaeizadeh, S., 2019. *An Ensemble of Bagged Decision Trees for Early Prediction of Sepsis*. Singapore, IEEE, pp. 1-4.

Fu, M., Yuan, J. & Bei, C., 2019. *Early Sepsis Prediction in ICU Trauma Patients with Using An Improved Cascade Deep Forest Model*. Beijing, China, 2019, IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), pp. 634-637.

Garcia-Gallo, J., Fonseca-Ruiz, N., Celi, L. & Duitama-Muñoz, J., 2019. *One-Year Mortality Prediction in ICU Patients with Diagnosis of Sepsis Driven by Population Similarities*. Athens, Greece, IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 480-484.

R. Gómez, N. García, G. Collantes, F. Ponce and P. Redon, "*Development of a Non-Invasive Procedure to Early Detect Neonatal Sepsis using HRV Monitoring and Machine Learning Algorithms*," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 2019, pp. 132-137..

Q. Gu, V. Prasad and T. Heldt, "*Characterizing Fluid Response and Sepsis Progression in Emergency Department Patients,*" *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 510-513..

Hsu, P.-Y. & Holtz, C., 2019. *A Comparison of Machine Learning Tools for Early Prediction of Sepsis from ICU Data*. Singapore, IEEE, pp. 1-4.

Hu, Y., Lee, V. C. & Tan, K., 2019. *An Application of Convolutional Neural Networks for the Early Detection of Late-onset Neonatal Sepsis*. Budapest, Hungary, International Joint Conference on Neural Networks (IJCNN), pp. 1-8.

R. Joshi, D. Kommers, L. Oosterwijk, L. Feijs, C. van Pul and P. Andriessen, "*Predicting Neonatal Sepsis Using Features of Heart Rate Variability, Respiratory Characteristics, and ECG-Derived Estimates of Infant Motion,*" in *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 681-692, March 2020.

C. Lin, J. Ivy and M. Chi, "*Multi-layer Facial Representation Learning for Early Prediction of Septic Shock,*" *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 840-849.

Q. Li, L. F. Huang, J. Zhong, L. Li, Q. Li and J. Hu, "*Data-driven Discovery of a Sepsis Patients Severity Prediction in the ICU via Pre-training BiLSTM Networks,*" *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019, pp. 668-673.

Q. Li, L. F. Huang, J. Zhong, L. Li, Q. Li and J. Hu, "*Data-driven Discovery of a Sepsis Patients Severity Prediction in the ICU via Pre-training BiLSTM Networks,*" 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 668-673.

R. Matthew, J. Chris, S. Salman, J. Russell, S.Supreeth, W. M Brandon, S. Ashish, N. Shamim and C. Gari D, 2019. *Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge* 2019. Singapore, IEEE.

M. I. Shah, J. Joseph, R. Kedia, S. Gupta and V. Sritharan, "*A Portable Colorimetric Reader for Early and Rapid Diagnosis of Sepsis,*" *2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT)*, Bethesda, MD, USA, 2019, pp. 83-86.

J. Singh, K. Oshiro, R. Krishnan, M. Sato, T. Ohkuma and N. Kato, "*Utilizing Informative Missingness for Early Prediction of Sepsis,*" *2019 Computing in Cardiology (CinC)*, Singapore, Singapore, 2019, pp. 1-4.

A. Tekin, M. Ulas and F. Uzun, "*Analysis of the Neonatal Sepsis Data Set with Data Mining Methods," 2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Ankara, Turkey, 2019, pp. 1-4.

T. Vicar, P. Novotna, J. Hejc, M. Ronzhina and R. Smisek, "*Sepsis Detection in Sparse Clinical Data Using Long Short-Term Memory Network with Dice Loss*," *2019 Computing in Cardiology (CinC)*, Singapore, Singapore, 2019, pp. Page 1-Page 4.

M. Vollmer, C. F. Luz, P. Sodmann, B. Sinha and S. Kuhn, "*Time-Specific Metalearners for the Early Prediction of Sepsis,*" 2019 Computing in Cardiology (CinC), Singapore, Singapore, 2019, pp. 1-4.