Sepsis Prediction using Machine Learning and Deep Learning Algorithms

MSc Research Project
Data Analytics

Terrance Thomas
Student ID: X18184928

School of Computing
National College of Ireland

Supervisor: Dr. Rashmi Gupta

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Terrance Thomas
Student ID:	X18184928
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Rashmi Gupta
Submission Due Date:	28/09/2020
Project Title:	Sepsis Prediction using Machine Learning and Deep Learning Algorithms
Word Count:	8313
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sepsis Prediction using Machine Learning and Deep Learning Algorithms

Terrance Thomas
X18184928

Abstract

This paper aims at presenting a method and comparing various machine learning (ML) and deep learning (DL) algorithms for predicting sepsis from clinical time-series data. Sepsis is among the most threatening condition that could occur during intensive care unit (ICU) treatment of a patient. Therefore, in this research, multiple models is applied to the data after they are cleaned and pre-processed to get the best results. In the research DL method such as long short-term memory (LSTM) and ML method such as decision tree (DT), random forest (RF), adaptive boosting (AdaBoost), k-nearest neighbours (KNN) and extreme gradient boosting (XGB) to the data set which is taken from 2 hospitals via an online challenge and has hourly data of over 40,000 patients. Data are processed in two separate ways and the best performing was used to apply all the models. Of which XGB performs the best with 0.98 of accuracy, 0.96 recall score, 0.98 f1 score, and 0.99 precision followed by RF as the second-best model.

1 Introduction

Sepsis is by far the most common complication of trauma patients; Sepsis is a life threatening condition that happens when tissue damage, organ failure, or death is caused due to the response of body to infection. About 5.8 million people die worldwide, about 40% of deaths took place while hospitalisation and of which 22% was caused by sepsis (Fu et al.; 2019). Sepsis is a major health problem with a significant health impact throughout the world. Sepsis is economically significant, costing about 6 billion pounds in the UK hospitals, costing over 20 billion USD in 2011 in the US which is around 5.2% of all the hospital costs in the US. Sepsis is defined as serious a "life-threatening dysfunction of the organ caused by a dysregulated response of the host to infection" (Li et al.; 2019; Mohamed et al.; 2020).

Delayed treatment every hour leads to an increase in the mortality of approximately 4% to 8%. Sepsis detection can reduce the occurrence of post-traumatic complications (Fu et al.; 2019; Mohamed et al.; 2020). It has been shown that early and targeted treatment enhances the sepsis outcomes. In the ICU, predictive models were used to improve care, and might be used for identification of patients at risk of becoming septic (Wang, Sun, Schroeder, Ameko, Moore and Barnes; 2018).

Sepsis detection is currently done on clinical evaluation and by analysing laboratory results or vital signs that is often non-sensitive (Amiri et al.; 2019). About, 18.8 million electronic health records (EHR) can be found in the ICU that provides a distinctive opportunity to get new insights for improved care. EHRs are now an important information source and the big drive for modern medicine with the advances in ML (Li et al.; 2019; Lin et al.; 2019). It can make more accurate, quick, and personalised decisions. But a very little study has concentrated on predicting the seriousness of sepsis in EHR patients. Many studies suggest that artificial intelligence models have enabled computer scientists to develop innovative decision support systems based on ICU. But most early ICU analytic on sepsis data focuses primarily on the risk of mortality prediction and antibiotics improve survival (Li et al.; 2019).

Figure 1: Sepsis sign and symptoms

Given the growing evolution of modern biomedical technology, preterm infants remain highly vulnerable to the infection, particularly those of very low birth weight(Hu et al.; 2019). Neonatal sepsis is a condition which during the initial 30 days presents signs of infection. Neonatal sepsis are of 3 types viz early onset, late onset, and very late onset.

The application of ML and neural networks (NN) techniques to clinical data have been studied by researchers since the early 1990s. As huge computing power is available now to perform ML and NN has been resurgent in recent years (Mohamed et al.; 2020). Recurrent neural networks (RNN) and particularly LSTM networks are presently a very powerful tool that is wildly used in many tasks, where information is a variable-length sequence, including signal classification, natural language processing (Vicar et al.; 2019).

In Section 2 all the related work in relation to sepsis has been discussed. In Section 3, the methodology followed for the project is explained in detail and the steps taken as well. In Section 4, the design of the project is given. In Section 5, the implementation of the models is explained. In Section 6, the evaluation matrix and model performance is compared. Section 7 explains the conclusion of the research and future work associated.

1.1 Motivation of this Research

Infections frequently occur in the lungs, belly, kidneys or bladder. Sepsis can starts with a small cut which becomes infected or by an infection which develops after surgery. Sepsis can sometimes develop in individuals who didn't even know they were infected[1]. Conventional ICU forecasting predictions are based on large population analysis but often provide statistically rigorous findings for an average patient, but are also very expensive, slow and also prone to bias in selection. Traditional methods tend to lack the accuracy

[1]https://medlineplus.gov/sepsis.html

necessary while at the individual level, as they show serious errors in patient information away from the average (Garcia-Gallo et al.; 2019).

About 5.8 million people die due to trauma of which 40% of deaths occurs while the patient is hospitalised of which almost 22% of deaths is caused by sepsis. Detection of sepsis reduces the hospital cost and increases the surviving chance of a patient (Fu et al.; 2019). Sepsis continues to progress to severe sepsis if there are indications of organ failure, such as difficulty in respiration, low or no urine output, impaired liver tests, and mental changes condition, in addition to signs of sepsis. Approximately all severely sepsis patients require treatment in an ICU[2].

Even though physicians have suggested new terminology for sepsis, there is still a problem of early diagnosis and treatment of sepsis and the limitations of early diagnosis are unknown. Survivors of sepsis can endure long-term physical as well as psychological problems that require more support for social and health services. Deaths mostly are of the elderly people, infants, and patients with weak immune systems. A study by the WHO in 2017 shows that infectious diseases led to 50% of the deaths in infants (Shah et al.; 2019).

Premature infants are at higher risk of having sepsis and multiple organ failures. Sepsis is a huge concern for public health, which highly contributes to infants mortality (Joshi et al.; 2020).

Since the people ho do not have a healthy immune systems like babies and aged persons are more susceptible to sepsis, it is very crucial to identify sepsis accurately so that such vulnerable people can be diagnosed and treated on time. ML has been used in many research to improve prediction of infection and diseases and hence directly helping to treat the patients early and reduce the cost of treatment and reduce the death rate.

1.2 Research Questions

The following are the research question that are answered in this research project:

- How different ML and DL algorithms can be used to predict sepsis accurately?
- To what extend can patient's medical record be used to enhance the performance various algorithms?

1.3 State of the Art

For the research, I have considered the research done by (Mohamed et al.; 2020) which was done in 2020. The research was done by comparing multiple models of ML and DL. This research will be following a similar pattern of working on multiple models and improve on the performance of few models used as in the State of the art (SOA) approach research and adding some additional models based on the literature review and comparing those papers in Section 2.1. Section 3.7 explains the models used for the research.

2 Related Work

Ribas et al. (2011) used relevance vector machines (RVM) in the research to provide an automatic ranking of death predictors. The database consists of severe sepsis patients

[2]https://www.sepsis.org/sepsis-basics/what-is-sepsis/

from June 2007 through December 2010 of 354 patients. RVM using an embedded feature importance selection process and is proved to be better in terms of accuracy than other well established methods with an AUC of 0.80.

Marshall et al. (2012) focuses on developing a separate conditional survival algorithm (DC-S) which has a classification element to predict patient outcome and their survival. The model DC-S has two main components; the conditional component that uses a tree of classification and the survival component that models the distribution of survival and gave an accuracy of 0.99.

Mani et al. (2013) aimed to develop a non-invasive prediction model from medical data and EMR for late onset neonatal sepsis. 299 infants information was considered for the research. Naive Bayes (NB), support vector machine (SVM), classifiers classification and regression trees (CART), KNN, lazy bayesian rules (LBR), RF, averaged one dependence estimators (AODE) and augmented naive Bayes (TAN) were applied to the data set. AODE had the best sensitivity with data set 1 at 0.88 while NB and RF had 0.95 and 0.94 respectively with data set 2.

Guillén et al. (2015) explored a new framework for severe sepsis prediction. There are few models such as logistic regression (LR), Logistic model trees (LMT), and SVM that uses the vital signs and laboratory results, or a mixture both. The SVM model correctly identified 65% of the patients of having severe sepsis.

Gunnarsdottir et al. (2016) built a generalised linear model (GLM) to predict sepsis in an ICU patients. The model was trained on 29 patient records and evaluated on a different test set of 8 patient records. Using demographic measures as features, an accuracy of 0.625 was obtained. The adding the physiological time series characteristic to the model increased the accuracy to 0.75.

Using vital signs and results of blood, LR, SVM, and LMT was applied to data which was used to predict sepsis in adult ICU patients. LMT produced superior performance, the specificity of LMT was 0.83 (Wang, Sun, Schroeder, Ameko, Moore and Barnes; 2018).

Thakur et al. (2018) developed an appliction to calculate the likelihood of sepsis. Non invasive variables were used to develop the prediction model which performed well in comparison with the invasive parameter prediction model. About 58,000 hospital admissions were used of which 38645 were adults while 7875 were infants. LR was used on both 3 invasive and also on 3 non invasive parameters for predicting. The AUROC was 0.777 and 0.824 for invasive and non invasive algorithms, and 0.830 and 0.824 for validation data set respectively.

A great learning method is proposed to improve the performance of the extreme learning machine kernel, known as the optimisation of a chaotic fruit fly. Feature selection was done using the RF before the classification model was constructed. A total of 42 patients with sepsis were utilized for the research. RF-CFOA-KELM achieved results of 0.8160 accuracy, 0.7766 MCC, 0.8957 sensitivity, and 0.6577 specificity (Wang, Wang, Weng, Wen, Chen and Wang; 2018).

Gómez et al. (2019) created a less invasive and cheap tool that uses heart rate variability (HRV) and ML algorithms. It is used to predict the sepsis risk in infants within the first 48 hours. 79 newborns aged between 36 and 41 weeks with gestational age were recruited. The research has implemented AdaBoost, LR, NB, bagged classification trees (BCT), RF, KNN, classification tree (CT), and SVM. The AdaBoost AUC is best with 0.94 followed by BCT and RF.

Wang et al. (2019) presents 2 methods from the clinical data for the early prediction of sepsis. One is an LSTM and the other is based on the XGB method. The data set consists of 40 characteristics such as values of Demographics, Vital Signs, and Laboratory. There is a total of 40336 patient records. Along the timeline, the currently missing data were filled with the last non-missing data from previous data when data was found missing. In the LSTM-based method, the remaining missing data after the preceding step was filled with value 0. The utility score is 0.267 for the LSTM-based method, and 0.392 for the XGB based method.

Alnsour et al. (2019) aims to predict death of patients with sepsis in hospitals. The data set was composed of 1,048,575 sepsis patients hospitalised between 2008 and 2012. The following models were applied LR, RT, BN, NN, SVM, chi-square automatic interaction detection (CHAID), and Quest. The results showed that CHAID had the best accuracy of 82.08% invalidation followed by Quest. In phase 2, The authors added the attributes of the health care provider to the data used to construct the models and set the type of each attribute accordingly. CHAID model had the best with an accuracy of 0.853.

Fu et al. (2019) suggest an early sepsis prediction algorithm which uses deep forest (DF) cascade model which is improved than the standard algorithm. 3125 patient data were used. The models were SVM, RF, LR, KNN, GB, XGB. The researcher's model gave a better AUROC of 0.80 than the traditional models. The accuracy of the model was 73% which was the best along with GB which gave an accuracy of 0.73 as well.

The paper focuses on assessing the impact and importance of 3 different patient information metric based on the similarity of a 1-year prediction of mortality when the patients were related to similar diagnosis of sepsis. 16219 admissions were obtained; newborn patients with 16,080 admissions were excluded from these. Finally, only hospital admissions where 15751 patients were selected for more than one day. The best LR algorithm gave an AUROC of 0.73 (Garcia-Gallo et al.; 2019).

A method has been developed that gives importance on identifying optimum HRV features in intensive pediatric care for early identification of sepsis. Four models of the classifiers viz KNN, DT, Linear Discriminant Analysis (LDA), and SVM were used. DT performed the best with an accuracy of 0.8636 (Amiri et al.; 2019).

A method has been developed which predicts sepsis early by observing the physiological features and by comparing them with general patterns of data. XGB model has been applied to that data set. The best model yielded a 0.8406 AUROC score (Singh et al.; 2019).

The paper developed a bagged decision trees (BDT) ensemble with a highly unbalanced misclassification cost to predict the sepsis for each sample of patient features. A 3 bagger classifier was applied to the data set using the training data set, and the features and hyperparameters were selected in an continues process till the best score was achieved. The accuracy of the model was 0.871 and 0.912 on the two data sets (Firoozabadi and Babaeizadeh; 2019).

Lin et al. (2019) suggests a general framework that uses the extracted temporary relationships and patterns through facial representations that has evolving emotional expressions based on the patient health conditions. 2D CNNs were used to obtain high level facial features from the images produced and then was fed the facial representations extracted into LSTM model to predict sepsis. The best accuracy is by the proposed model dyn+ img.2D CNN.LSTM + sta of 0.9037.

The research is on neonatal sepsis data set analysis using DM techniques. 13 variables are present in the data set. KNN and NB algorithms are applied to the data set. KNN algorithm gave an 0.9453 accuracy while NB algorithm produced a 0.9375 accuracy (Tekin et al.; 2019).

An attention based algorithm in sepsis prediction that gives more information on the amount of contribution to the final prediction of each of the medical measurements is explored. More than 53,000 adult patient's records were used. In total there are 11,791 patients used for thestudy. A total of 39 features were selected. The research is utilising RNN as specific bidirectional RNNs. The model gives a precision of 0.75 (Baghaei and Rahimi; 2019).

Hu et al. (2019) focuses on creating an application for preterm newborns which predicts sepsis at an early stage using CNN. The data set is composed of 146 newborns. Data chunks was used which are nothing but sequences of data cut and transformed to a certain length. The study used a 14 layer deep CNN which an AUC 0.79 and precision is 0.76.

Bidirectional LSTM is used for predicting the seriousness of sepsis in ICU patients. For the prediction 48 features were used from the data set. Bidirectional LSTM is made of the sub-LSTM, which moves forward as well as backward. The proposed model achieves an F1 score of 0.9472 (Li et al.; 2019).

The study is on sepsis patient identification based on EMR in the emergency department. DT, Discriminant Analysis (DA), LR, KNN, Ensemble Classification (EC), SVM, and NN are the techniques used in ML. Also includes a novel rule-based genetic-algorithm-optimized system. The data set consists of 912 sepsis and 975 non sepsis patients. The NN model yielded the best results. 65 hidden neurons with 0.9208 sensitivity and 0.9233 specificity were the best performing NN model (Mohamed et al.; 2020).

The study focuses on comparing the performance of many major ML algorithms to identify patients with sepsis in the Emergency Department (ED). The following models were applied to compare DT, DA, LR, KNN, Ensemble Classification, SVM and NN are the techniques used in ML. A novel, genetic algorithm optimised rule based system developed by the authors is also used. The data set had 1,887 unique cases. The NN model yielded high performance with sensitivity of 0.9208, specificity of 0.9233, PPV of 0.9178, and accuracy of 0.9221 (Mohamed et al.; 2020).

The research presents a model for overcoming these deficiencies using a DL approach to a diverse multicenter set of data. The research had 3 main approaches: a GB Vital system based on vital sign characteristics, a non sequential MLP model of thousands of characteristics, including those used for the GB Vital model; and a sequential CNN+LSTM model with an equal number of functions. GB Vital model had decent performance giving an AUROC of 0.786 3 hours before sepsis onset. CNN+LSTM model achieved an AUROC of 0.856 when evaluated for 3h (Lauritsen et al.; 2020).

Joshi et al. (2020) aims on predicting sepsis in premature infants. 22 features of the infants was used for the research of 49 infants. The data set was used NB classifier. The root mean square of the successive differences (RMSSD) was 66% and the average response was 61%.

2.1 Approach comparison

This section compares the different important papers and including the models applied, aim and objective, the data set. These are the primary papers considered while applying the model in the research which contains the Author and title, aim and objectives, models which were applied, data set used and the final outcomes as shown in Table 1.

Table 1: Important Paper Comparison

Author(s) and Title	**Aims and objective**	**Models Applied**	**Data set**	**Findings relevant to the review**
Mani et al. (2013) - "Medical decision support using machine learning for early detection of late-onset neonatal sepsis"	"To develop non-invasive predictive models for late-onset neonatal sepsis from off-the-shelf medical data and electronic medical records"	NB, SVM, RF, KNN, CART, LBR, AODE, and TAN	Infants admitted to the NICU for 18 months from 1 January 2006 with 299 infants	AODE had sensitivity of 0.88 with data set 1 and with data set NB and RF had 0.95 and 0.94 sensitivity respectively.
Wang et al. (2019) - "Prediction of Sepsis from Clinical Data Using Long Short-Term Memory and eXtreme Gradient Boosting"	"To develop an objective and efficient computer-aided tool for early detection of sepsis"	LSTM and XGB	Data of two independent hospitals were used. A total of 40,336 records.	The utility score is 0.267 for the LSTM-based method, and 0.392 for the XGB based method.
Amiri et al. (2019) - "Identifying Optimal Features from Heart Rate Variability for Early Detection of Sepsis in Pediatric Intensive Care"	"Early diagnosis of sepsis before clinical signs are developed to give physicians enough time for antibiotic therapy of these patients"	SVM, LDA, KNN, and DT	Data was of two PICU hospital and the patients were between 2016 and 2018 and had about 500 records	DT had the best accuracy of 0.8636.
Singh et al. (2019) - "Utilizing Informative Missingness for Early Prediction of Sepsis"	"To predict the occurrence of sepsis early by studying the missingness of physiological variables and using it with the overall trends in data"	XGB with multiple variations	Data comprised of three distinct hospitals in the United States with ICU stay records of 40,336 patients	Best Model of XGB got a 0.8406 AUROC score.
Tekin et al. (2019) - "Analysis of the Neonatal Sepsis Data Set with Data Mining Methods"	"To address the relationship between data mining methods and health services."	KNN and NB	Data set used is from Fırat University Hospital's Pediatrics Department collected between May 2013 and January 2014	KNN algorithm made 0.9453 accuracy.
Gómez et al. (2019) - "Development of a Non-Invasive Procedure to Early Detect Neonatal Sepsis using HRV Monitoring and Machine Learning Algorithms"	"To develop a minimally invasive and cost-effective tool, based on HRV monitoring and ML algorithms, to predict sepsis risk in neonates within the first 48 hours of life."	AdaBoost, RF, BCT, CT, LR, NB, KNN, and SVM	79 infants between 36 and 41 weeks data was recorded	AdaBoost AUC is best with 0.94.
Mohamed et al. (2020) - "Electronic-Medical-Record-Based Identification of Sepsis Patients in Emergency Department: A Machine Learning Perspective"	"To compare the performance of several major machine learning techniques to identify emergency department sepsis patients during their first 6 hours of care."	DT, DA, LR, KNN, EC, SVM, NN, and genetic algorithm optimised system	1887 patients from the ED of Detroit Medical Center in Michigan, USA	NN model achieved 0.9208 sensitivity, 0.9233 specificity, and 0.9178 PPV.

3 Methodology

For this research project, Knowledge discovery in databases (KDD) was used. KDD is an incremental method where measurements for assessment can be improved, mining can be refined, the latest data is integrated and converted to obtain different and more appropriate results[3].

KDD has 5 important phases as shown in figure 2. Phase one is called as Data selection where the data set for the research or study is selected. The second phase is prepossessing of the data where cleaning data, deleting unnecessary data, imputing missing data. The third phase is called transformation which transforms data, as per the algorithm or model. This can also include converting data from a complicated format into one that is simple. Phase four is called Data Mining, this is where the proposed ML algorithm is implemented. Upon completion of the training, the model runs on test data. The fifth and final phase is termed as interpretation. In this phase, knowledge is acquired and insights learned.

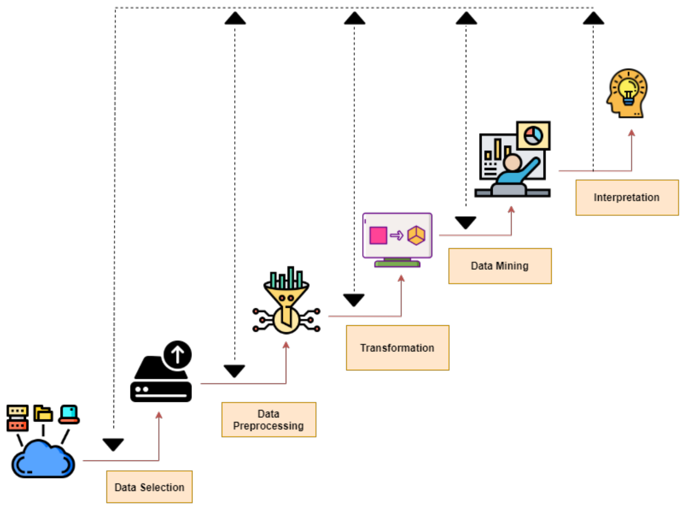

Figure 2: Knowledge Discovery in Databases

3.1 Data set

Figure 3 shows us the data set variables. Data used for this research is from from ICU patients of two separate hospitals. The data set was part of a challenge and openly available on https:/physionet.org/content/challenge-2019/1.0.0/.

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 3: Data set Variables

[3]https://www.geeksforgeeks.org/kdd-process-in-data-mining/

Each file has the similar header and each row which represents the value of a single hour of data. Data available for patients are Demographics, Vital Signs, and Laboratory values. The data set contains 20,336 and 20,000 patient records. Each table column provides a sequence of measurements over time, in which the column header explains the observation. There are 40 variables that depend on time. SepsisLabel, indicates sepsis where 1 shows sepsis and 0 shows no sepsis.

3.2 Exploratory Data Analysis

In this section, the data is explored and checked before data pre-processing is done. When the records are merged the hourly data sums up to 1,552,210 rows. As seen in figure 4 only 2% of the total patient records shows sign of sepsis. Of 40,336 patients in 2,932 patients had sepsis. This shows how the data set is highly imbalanced for the predicting/dependent variable.

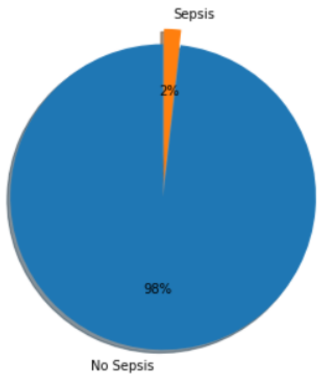

Figure 4: Sepsis records

Figure 5 shows the trend of how long the patients were admitted to the hospital. It can be seen that post 60 hours the number of patients is very less.

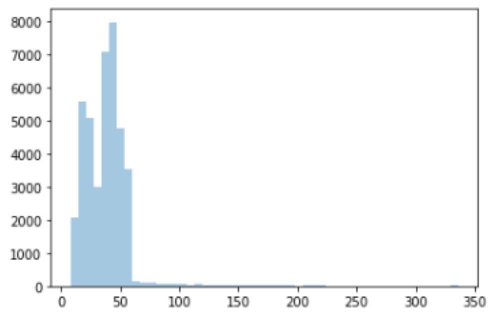

Figure 5: Hourly graph of patients admitted

3.3 Data Pre-processing

Pre-processing of data is performed to make sure no inconsistencies were found in the data. Irrelevant and redundant present or noisy and unreliable information is removed from the data set. First, the most important thing was to check how much of the data is missing from the entire data set.

Figure 6 provides an overall information on the missing data. EtCO2, HCO3, BaseExcess, FiO2, pH, PaCO2, SaO2, BUN, AST, Alkalinephos, Calcium, Creatinine, Chloride, Bilirubin_direct, Glucose, Magnesium, Lactate, Phosphate, Potassium, Bilirubin_total, Hct, TroponinI, Hgb, PTT, WBC, Fibrinogen and Platelets had more than 80% of the data missing. Temp had more than 60% of the data missing while Unit1 and Unit2 had more than 40% of the data missing. Features having more than 40% missing data are dropped.

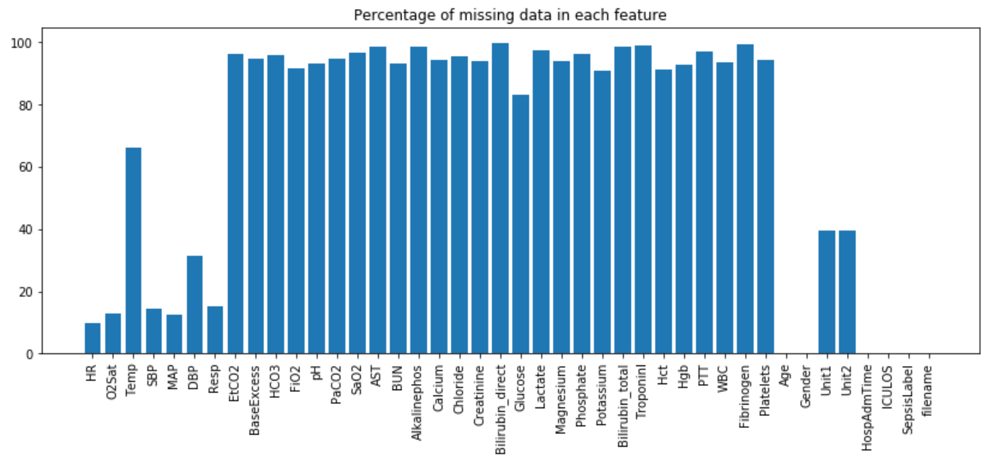

Figure 6: Missing data from the data set

Figure 7 shows the correlation of the variables with each other. The correlation matrix shows the degree to which a pair of variables are linearly related. The variables can be correlated positively or negatively. If they are positively correlated then the color is towards maroon. Similarly, if the variables are negatively correlated then they are marked with blue color. SBP and MAP are strongly correlated positively while SBP and DBP are slightly correlated positively. All the other variables are loosely correlated and the correlation value is almost zero.

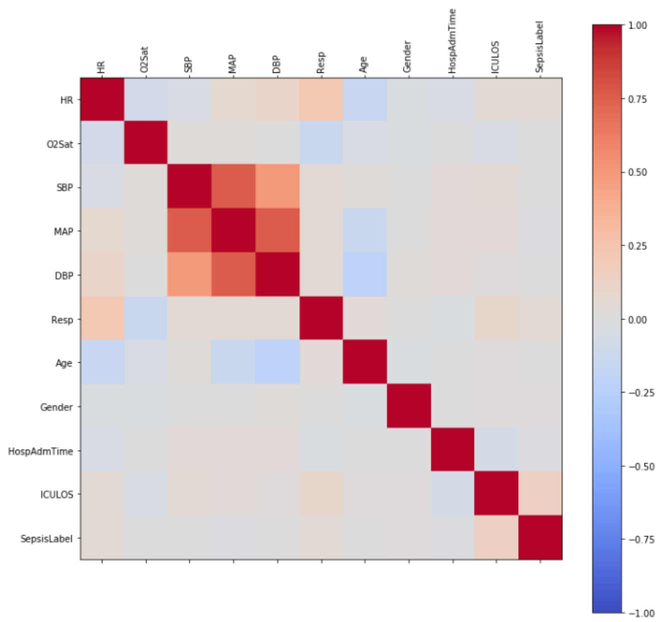

Figure 7: Correlation matrix

The next step is data imputing. This is the process of substituting missing data with replaced values. For this research, median substitution is used for data imputation. This involves to calculate the median value of the non missing values in the column and then substitute in place of the missing values. Some of the benefits of this type of imputation are that it is easy and fast, working well with numerical data sets[4].

3.4 Feature Extraction and Encoding

Feature Extraction helps to reduce the features in a data set by creating new features from existing features. This makes it possible to create a summarized version of the original features from a combination of the original set[5]. In the research, there were 11 features left post data cleaning and data imputation. Of which 10 being independent feature while 1 being the dependent feature. SBP stands for Systolic blood pressure which is upper number while measuring Blood pressure while DBP stands for Diastolic blood pressure, It is the lower number while measuring Blood pressure.

In this research, we combined both the features to get blood pressure as low, normal, elevated, high, and missing (which does not fall in any categories). This missing value had to be filled with median imputation. Other features/variables which were changed were Age, O2Sat, Respiration, MAP, and heart rate where the values were converted to a range and if there were any missing values then it had to be imputed as well. Also, Values of the data set were converted into fixed values based on the medical information also called as encoding.

3.5 Under Sampling

If a data class is the over-represented majority class, this can be used to balance it with the minority class under-sampling. Under-sampling is used when the amount of data collected is sufficient. Trying to reduce the bias associated with imbalanced data classes under-sampling. Overall the majority class works best for large data sets under-sampling. In the research, NearMiss library is used for under-sampling.

3.6 Data Splitting

There are 2 different data splitting in the research. The first one with the data into train and test data set of the data after feature extraction and under-sampling done. The second data splitting was to the data without applying feature extraction but using under-sampling. In both, the scenario the actual data is split into train data which is 75% of the data used for training while the test data is the remaining 25% of the data.

3.7 Models

- **Long Short Term Memory**: It is a kind of artificial NN designed for recognizing patterns in data sequences. LSTM help preserves the error that can be over time and layers propagated backward. By keeping a more consistent error, they allow recurrent networks to continue learning in many time steps[6]. As shown in the

[4]https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

[5]https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be

[6]https://pathmind.com/wiki/lstm

research of (Wang et al.; 2019) LSTM works well and gives good results.

- **Random Forest**: It is an ensemble method that integrates more than one method of the same or different sort for object classification. RF classifier helps to create a set of DT. It then collects the votes from the different DT for the final class of the test object to be decided[7]. As shown in the research (Mani et al.; 2013) RF has a decent performance in 2 different data sets.

- **Decision Tree**: DT is constructed through an algorithmic approach that finds ways to split a data set based on various rules or conditions. DT is used widely and is a non parametrically supervised method of learning which is used for classification and regression. The aim is to create a model which predicts the value of the dependent or target variable by learning simple rules of decision learnt from the data characteristics[8]. Amiri et al. (2019) shows DT performs well for sepsis detection.

- **Extreme Gradient Boosting**: It utilises the mentioned techniques with boosting and is wrapped in a library that is easy to use.Some of XGBoost's major advantages are that its highly scalable / parallel, quick to implement and usually outperforms other algorithms[9] also shown is research done by (Wang et al.; 2019; Singh et al.; 2019).

- **Adaptive Boosting**: It is a meta estimator which starts by fitting the classifier to the original data set and then later fits additional duplicates of the classifier to the existing data set but then adjusts the weights of the wrongly classified instances so that the subsequent classifiers focus more on difficult cases [10]. Sepsis detection is done with good results as shown by (Gómez et al.; 2019) in their research.

- **K-Nearest neighbors**: It uses feature resemblance o similarity to predict the values of new data points, that the new data point will be given a value based on how closely it fits the training points[11]. KNN performs well with text data and was used by (Tekin et al.; 2019) for the research.

4 Design Specification

As mentioned in Section 3, the research follows a KDD approach for the execution. The local system of processor i7 8th generation with 1TB hard drive, 16 GB RAM, and 8 GB of the graphic card was used to execute the project code. The code ran on python version 3.7.3 on the jupyter notebook. Multiple python packages had to be imported for the execution of the project.

Figure 8 shows the detailed full process flow and design of the research approach. The downloaded data from the website consists of 2 separate data set viz training set A and training set B. Both the data set are loaded in data frames in python and them combined. An alternate way was to download the 2 data set and combine the 2 data set on the excel

[7]https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1

[8]https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/

[9]https://www.aitimejournal.com/@jonathan.hirko/intro-to-classification-and-feature-selection-with-xgboost

[10]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[11]https://www.tutorialspoint.com/machine$_{l}earning_{w}ith_{p}ython/machine_{l}earning_{w}ith_{p}ython_{k}nn_{a}lgorithm_{f}inding_{n}e$

folder and use that to load in python. Now the data pre-processing part starts as shown in Section 3.3.

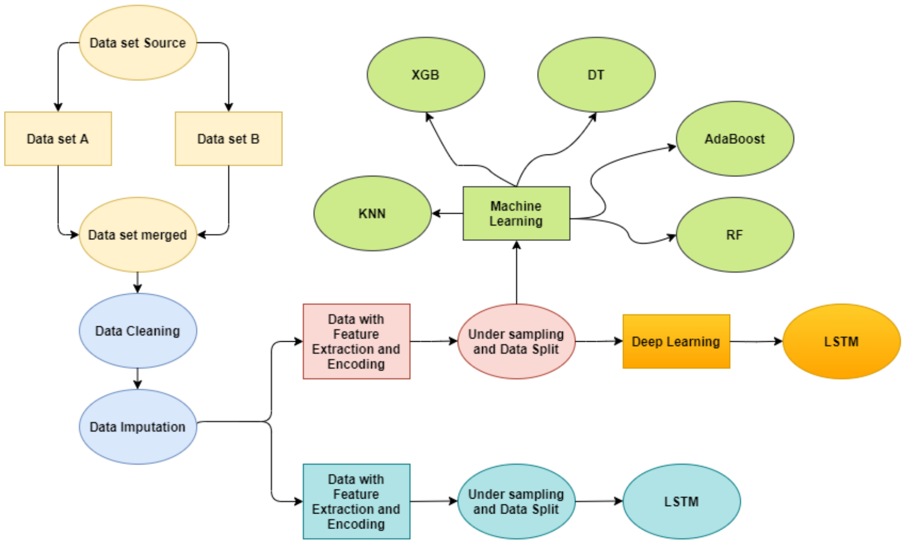

Figure 8: Design Architecture

After the pre-processing, there are 2 different scenarios. Scenario 1 is when feature extraction and encoding are applied to the data set as mentioned in section 3.4. Once the imputation was done data under-sampled to compensate for the data imbalance. To overcome the imbalanced data set, under-sampling was applied to get 55,832 rows which involve making the data set of equal sepsis and non-sepsis case. Now the data was divided into the train (75% of data) and test (25%) of the remaining data. LSTM is applied to this data set which had an accuracy less than the data set which did no undergo Feature Extraction and Encoding. Hence, it was not considered for other models.

Scenario 2 is when no feature extraction or encoding is applied to the data set. Hence directly under sampling is done. Then data is split into 2 parts viz train and test data set. Now multiple models are applied on this data set including LSTM, RF, Adaboost, XGB, KNN, and DT.

5 Implementation

5.1 Deep Learning

- **Long Short-Term Memory**: The data had to be converted from a 2 dimension to a 3 dimension in order to feed it into for LSTM. The input shape was (1,9) for data set with feature extraction and encoding while it was (1,10) for the data set without feature extraction and encoding. The input shape is by the number of input variables to the output variables.

  Figure 9 shows the LSTM design used for the research. There is 2 layer of LSTM 256 where 256 is the hidden nodes or the neurons. This is followed by a layer with ReLU and a dense layer with Sigmoid. The loss is checked using mean square error and adam optimiser is used. The batch size is kept as 150 and the epoch is set to 50 and 45 for the 2 different training and testing data. The data with feature extraction and encoding gave a test accuracy of 0.9181, F1 score of 0.91, precision of 0.97, and recall value of 0.85 while the data set without feature engineering and encoding gave a test accuracy of 0.97, F1 score of 0.97, precision of 0.98, and recall value of 0.96. For LSTM, the last epoch value is considered in the research.

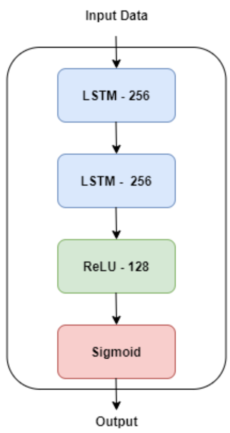

Figure 9: LSTM design

5.2 Machine Learning

- **Random Forest**: The maximum depth is was set to 15 with n estimators kept as 100 while the class weight is kept to a balanced subsample. The accuracy of the model is 0.98, f1 score is 0.98, precision is 0.99 and recall is 0.96.

- **Decision Tree**: DT works by data splitting which means into binary values until the prediction can be done. Max depth is kept as 5 and the random state is kept as 0. The accuracy of the model is 0.97, f1 score is 0.97, precision is 0.99 and recall is 0.95.

- **Extreme Gradient Boosting**: XGB parameter min child weight is set to 1,5 and 10 which specifies the minimum amount of the weights required for all observations. Gamma is set to 0.5, 1, 1.5, 2, and 5. The subsample is set to 0.6, 0.8, and 1.0. Colsample bytree is set to 0.6, 0.8, and 1.0. Finally, max depth is set to 1,2,3,4 and 5 which can be used to control over-fitting. The accuracy of the model is 0.98, f1 score is 0.98, precision is 0.99 and recall is 0.96.

- **Adaptive Boosting**: Adaboost assists in combining many weak classifiers to a single powerful classifier. It works by placing more weight on instances that are difficult to classify and less on those already handled well. The accuracy of the model is 0.97, f1 score is 0.97, precision is 0.99 and recall is 0.96.

- **K-Nearest neighbors**: Works on neighbors vote and for the research, all the parameters were set to default values. The accuracy of the model is 0.97, f1 score is 0.97, precision is 0.99 and recall is 0.95.

6 Evaluation

6.1 Evaluation matrix

For this research Accuracy, F1 score, Precision, and Recall have been considered. These 4 values can be derived from a confusion matrix. The are 4 important factors or values in the confusion matrix:

- **True Positive (TP)**: It is the case in which model predicted as YES and the actual output were also is a YES.

- **True Negative (TN)**: Is a case when the model predicts negative class correctly. In simple words, it is the cases where model predicted as NO and the actual output is also a NO.

- **False Positive (FP)**: Is a case when the model wrongly predicts a positive class. In other words, where the model forecast as YES, and the actual output is a NO.

- **False Negative (FN)**: Results when the negative class is incorrectly predicted by the model. In other words, where the model forecast NO and the actual output was YES.

6.1.1 Accuracy

It is the number of predictions that are correct to total number of input samples[12]. The formula is:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

6.1.2 F1 Score

It tells how accurate the classifier is and how robust the model is. The formula is:

$$F1score = 2 \times \frac{1}{\frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}}} \tag{2}$$

6.1.3 Precision

It is ratio of correct positive outcomes to the number of positive outcomes predicted by the qualifier. Precision attempts to respond to what percentage of positive identifiers was actually correct. The formula is:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

[12]https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

6.1.4 Recall

That is ratio of correct positive results to the number of all relevant samples. Recall tries to answer which proportion of positive actual was correctly identified. The formula is:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

6.2 Experiment

6.2.1 LSTM on data with and without feature extraction and encoding

The main reason for this experiment is to determine which data is to be selected. As mentioned in section 4 there are 2 scenarios for the data set. This experiment is of scenario 1.

Figure 10 shows that the test accuracy for data with feature engineering and encoding is between 0.90 and 0.92 while the test accuracy of the data without feature engineering lies between 0.96 to 0.98 other than the drop in one occasion to 0.95. Similarly, when compared F1 score, the value of data with feature extraction and encoding lies in between 0.89 and 0.91 to 0.94 and 0.98. When compared to the precision, the value of data with feature extraction and encoding lies in between 0.91 and 0.98 while data without feature extraction and encoding lie between 0.97 and 0.99. The Recall score of data with feature extraction and encoding is between 0.83 to 0.88 whereas recall of data without feature extraction and encoding is between 0.92 and 0.97.

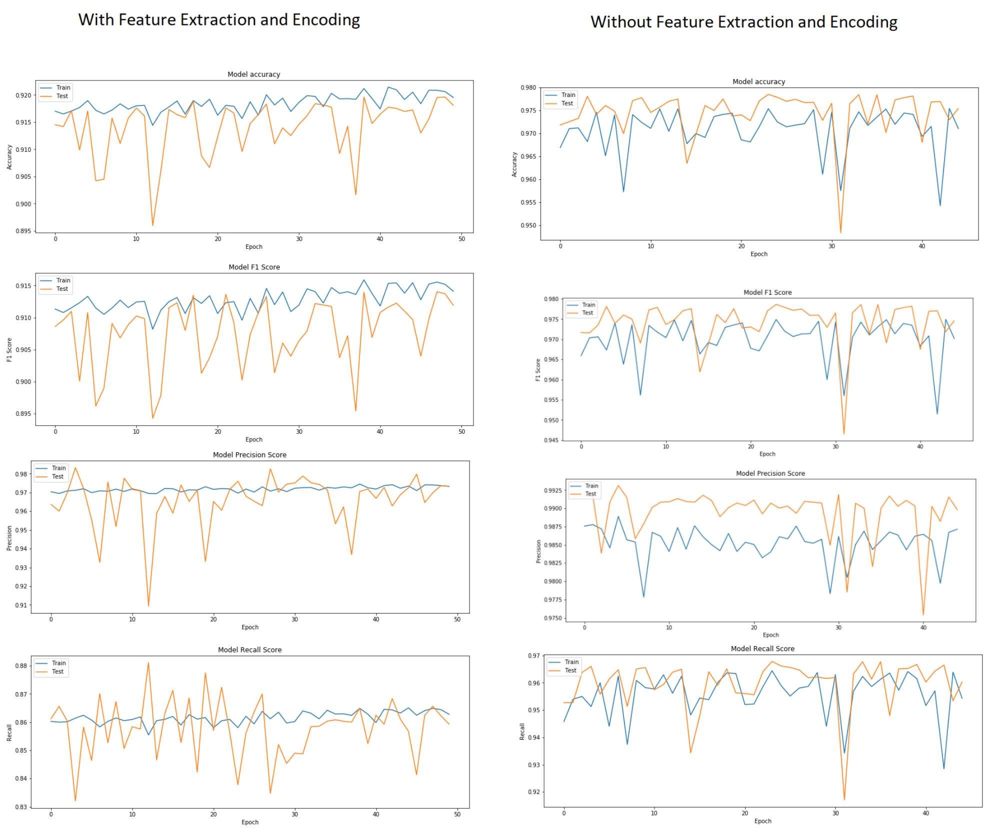

Figure 10: Long Short-Term Memory comparison

It is seen that data without feature extraction and encoding has better performance for the LSTM model applied. Even the model loss is less when compared. Based on the results only data without feature extraction and encoding are considered further for the experiment where different models are applied. The data feature which was extracted did not perform better and then the further missing values meant again imputation which is not considered good as the data was already imputed earlier.

6.2.2 Deep Learning and Machine Learning models

In this experiment, DL models and ML models are applied to the data set which is cleaned and imputed. The objective of this experiment is to check the performance of various models. Figure 16 shows the table of the different model's performance using 4 parameters viz Accuracy, F1 score, Precision, and Recall.

Table 2 shows complete performance of all the models applied. If accuracy is considered, XGB has the best with 0.98 followed by RF of 0.98. In terms of F1 score, again XGB has the best performance of 0.98 followed by RF with 0.98. XGB and RF perform the best in precision with 0.99 and for Recall XGB outperforms other models with 0.96 followed by RF with 0.96.

Table 2: Model comparison

Model	LSTM	RF	DT	XGB	AdaBoost	KNN
Accuracy	0.97	0.98	0.97	**0.98**	0.97	0.97
F1 Score	0.97	0.98	0.97	**0.98**	0.97	0.97
Precision	0.98	**0.99**	0.99	**0.99**	0.99	0.99
Recall	0.96	0.96	0.95	**0.96**	0.96	0.95

The SOA had its performance calculated in sensitivity, specificity, and PPV. Sensitivity is also called Recall and PPV is called precision. Since for the research, both Recall and Precision were calculated, will use those parameters to compare the research model against the SOA model.

From table 3 it can be seen that, the best model of the research which is XGB outperformed the best model of the SOA model which was a NN model in both Precision and Recall.

Table 3: SOA model vs Research model

	SOA - NN	Research - XGB
Precision	0.91	**0.99**
Recall	0.92	**0.96**

6.3 Discussion

XGB outperforms all the other models applied to the data set followed by RF. The performance can be improved by combination of multiple models. The research explained how the various ML and DL method performs on the data set and it shows ML models gives a better result than LSTM. For the data set used a lot of data was missing and hence 30 features or variables had to be dropped. The remaining data was used by data imputation and making the best use of the data available. However, if more data was available, it could be used to check which variables have the highest influence and by removing variables how the results alter. It might lead to better performance from other models as well. Feature selection could be used if all the variable data was available to improve the result. Also, different data imputation method could be used to see if that affects the result and how would the models perform on such data.

7 Conclusion and Future Work

Predicting sepsis accurately was one of the objectives of the research and the experiment shows how multiple models do indeed help in predicting sepsis from the data. As discussed in section 6 XGB and RF are the two models that were the best of all the models applied with the accuracy of 0.98 and 0.98 respectively. Similarly, if we consider F1 score, Precision and Recall as well as XGB and RF outperform all the other models. The second objective of the research is to what extent can the patient's medical record or information be used to enhance the performance of the model. From this research, it can be seen that of the 40 variables that were measured for every patient only 10 were used. This was because the rest of the data had a lot of missing values. But of the 10 variables which had data with little imputation, the prediction could be made which is better than the paper reviewed.

However, considering that the research is done on medical records and even 0.98 - 0.99 accuracy may seem good when applied to large people the 0.01 or 0.02 where the model fails is significant and future work involves using a better model or combination of the model to check if the performance improves. The accuracy may improve if the data set did not have a high percentage of missing values and provided could use the other variables which were dropped. Then using variable importance or feature selection the best variables can be selected to improve the existing model or the new model.

References

Alnsour, Y., Hadidi, R. and Singh, N. (2019). Using data analytics to predict hospital mortality in sepsis patients, *International Journal of Healthcare Information Systems and Informatics (IJHISI* **14**(3): 40–57.

Amiri, P., Derakhshan, A., Gharib, B., Liu, Y. H. and Mirzaaghayan, M. (2019). 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)*, pp. 1425–1428.

Baghaei, K. T. and Rahimi, S. (2019). Sepsis prediction: An attention-based interpretable approach, *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6.

Firoozabadi, R. and Babaeizadeh, S. (2019). An ensemble of bagged decision trees for early prediction of sepsis, *2019 Computing in Cardiology (CinC)*, pp. Page 1–Page 4.

Fu, M., Yuan, J. and Bei, C. (2019). Early sepsis prediction in icu trauma patients with using an improved cascade deep forest model, *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 634–637.

Garcia-Gallo, J. E., Fonseca-Ruiz, N. J., Celi, L. A. and Duitama-Muñoz, J. F. (2019). One-year mortality prediction in icu patients with diagnosis of sepsis driven by population similarities, *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 480–484.

Guillén, J., Liu, J., Furr, M., Wang, T., Strong, S., Moore, C. C., Flower, A. and Barnes, L. E. (2015). Predictive models for severe sepsis in adult icu patients, *2015 Systems and Information Engineering Design Symposium*, pp. 182–187.

Gunnarsdottir, K., Sadashivaiah, V., Kerr, M., Santaniello, S. and Sarma, S. V. (2016). Using demographic and time series physiological features to classify sepsis in the intensive care unit, *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 778–782.

Gómez, R., García, N., Collantes, G., Ponce, F. and Redon, P. (2019). 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)*, pp. 132–137.

Hu, Y., Lee, V. C. S. and Tan, K. (2019). An application of convolutional neural networks for the early detection of late-onset neonatal sepsis, *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Joshi, R., Kommers, D., Oosterwijk, L., Feijs, L., van Pul, C. and Andriessen, P. (2020). Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ecg-derived estimates of infant motion, *IEEE Journal of Biomedical and Health Informatics* **24**(3): 681–692.

Lauritsen, S. M., Kalør, M. E., Kongsgaard, E. L., Lauritsen, K. M., Jørgensen, M. J., Lange, J. and Thiesson, B. (2020). Early detection of sepsis utilizing deep learning on electronic health record event sequences, *Artificial Intelligence in Medicine* **104**: 101820.

Li, Q., Huang, L. F., Zhong, J., Li, L., Li, Q. and Hu, J. (2019). 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)*, pp. 668–673.

Lin, C., Ivy, J. and Chi, M. (2019). Multi-layer facial representation learning for early prediction of septic shock, *2019 IEEE International Conference on Big Data (Big Data)*, pp. 840–849.

Mani, S., Ozdas, A., Aliferis, C., Varol, H. A., Chen, Q., Carnevale, R., Chen, Y., Romano-Keeler, J., Nian, H. and Weitkamp, J.-H. (2013). Medical decision support using machine learning for early detection of late-onset neonatal sepsis, *Journal of the American Medical Informatics Association* **21**(2): 326–336.

Marshall, A. H., Payne, K., Cairns, K. J., Craig, S. and McCall, E. (2012). Modelling the development of late onset sepsis and length of stay using discrete conditional survival models with a classification tree component, *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 1–6.

Mohamed, A., Ying, H. and Sherwin, R. (2020). Electronic-medical-record-based identification of sepsis patients in emergency department: A machine learning perspective, *2020 International Conference on Contemporary Computing and Applications (IC3A)*, pp. 336–340.

Ribas, V. J., López, J. C., Ruiz-Sanmartín, A., Ruiz-Rodríguez, J. C., Rello, J., Wojdel, A. and Vellido, A. (2011). Severe sepsis mortality prediction with relevance vector machines, *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 100–103.

Shah, M. I., Joseph, J., Kedia, R., Gupta, S. and Sritharan, V. (2019). A portable colorimetric reader for early and rapid diagnosis of sepsis, *2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT)*, pp. 83–86.

Singh, J., Oshiro, K., Krishnan, R., Sato, M., Ohkuma, T. and Kato, N. (2019). Utilizing informative missingness for early prediction of sepsis, *2019 Computing in Cardiology (CinC)*, pp. 1–4.

Tekin, A., Ulas, M. and Uzun, F. (2019). Analysis of the neonatal sepsis data set with data mining methods, *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pp. 1–4.

Thakur, J., Pahuja, S. K. and Pahuja, R. (2018). Neonatal sepsis prediction model for resource-poor developing countries, *2018 2nd International Conference on Electronics, Materials Engineering Nano-Technology (IEMENTech)*, pp. 1–5.

Vicar, T., Novotna, P., Hejc, J., Ronzhina, M. and Smisek, R. (2019). Sepsis detection in sparse clinical data using long short-term memory network with dice loss, *2019 Computing in Cardiology (CinC)*, pp. Page 1–Page 4.

Wang, R. Z., Sun, C. H., Schroeder, P. H., Ameko, M. K., Moore, C. C. and Barnes, L. E. (2018). Predictive models of sepsis in adult icu patients, *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 390–391.

Wang, X., Wang, Z., Weng, J., Wen, C., Chen, H. and Wang, X. (2018). A new effective machine learning framework for sepsis diagnosis, *IEEE Access* **6**: 48300–48310.

Wang, Y., Xiao, B., Bi, X., Li, W., Zhang, J. and Ma, X. (2019). Prediction of sepsis from clinical data using long short-term memory and extreme gradient boosting, *2019 Computing in Cardiology (CinC)*, pp. Page 1–Page 4.