Early Sepsis Diagnosis using Machine Learning

Matjaž Muc
Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana
email: mm1706@student.uni-lj.si

***Abstract*—Sepsis is a potentially life-threatening condition that occurs as a response to body's response to infection. It might cause tissue damage, organ failure or can even cause death. Internationally, an estimated 30 million people have developed Sepsis and more than 6 million people die from it each year. The condition is among the hardest medical conditions to diagnose. Hence coming up with a model that might be able to solve this issue becomes imminent and important.**
In this paper we aim to analyze and evaluate several algorithms in hopes to answer some questions we face.

***Index Terms*—Overview of sepsis symptoms and reason on why sepsis is hard to diagnose, solutions to common problems in the medical data, evaluation and comparison of several models and strategies.**

I. Introduction

Early detection of sepsis is extremely important for improving sepsis outcomes. Each hour of delayed treatment can increase mortality rate by 4-8%, which typically ranges from 27-42%. The condition is known to be one of the hardest medical conditions to diagnose.

Immediate challenges we encounter during the analysis come from the nature of the dataset which consist of real medical data and is therefor far from perfect. The dataset is heavily skewed, containing only 2% of sepsis positive patients. Dataset contains about 50% of features that have more than 90% of missing values. After exploring feature relations we realize that the boarder between the two classes is not well defined, which is to be expected since the condition is hard to diagnose. Even more in most patients it is not very different from non-septic infection. We should also realise we are classifying between patients in Intensive care units (ICUs), therefor their health condition is severe regardless of them having sepsis or not. We attempt to solve the imbalance problem by using strategies presented in paper [1]. In order to solve the in-completeness of data, we pick features that differ between classes and impute the missing values for them. In this paper we will explore and evaluate performance of a logistic regression classifier, random forest model, isolation forest, neural network.

In this paper we start by taking a look at the general data information such as its origin, reliability, and structure. We proceed with an analysis of the data, our discoveries and challenges we face are presented. Proceeding is the introduction of solutions to missing values problem and imbalance problem. Later results of 4 distinct algorithms and techniques are presented. Finally we present paper discoveries and takeaways in the results and conclusions section.

December 24, 2021

II. Materials and Methods

A. Related work

1) Md. Mohaimenul Islam, Tahmina Nasrin, Bruno Andreas Walther, Chieh-Chen Wu, Hsuan-Chia Yang, Yu-Chuan Li,Prediction of sepsis patients using machine learning approach: A meta-analysis,Computer Methods and Programs in Biomedicine,Volume 170, 2019, Pages 1-9,ISSN 0169-2607.

2) X. Wang, Z. Wang, J. Weng, C. Wen, H. Chen and X. Wang, A New Effective Machine Learning Framework for Sepsis Diagnosis, in IEEE Access, vol. 6, pp. 48300-48310, 2018, doi: 10.1109/ACCESS.2018.2867728.

3) Silvia Cateni, Valentina Colla, Marco Vannucci,A method for resampling imbalanced datasets in binary classification tasks for real-world problems,Neurocomputing,Volume 135,2014,Pages 32-41,ISSN 0925-2312.

B. Data Overview

The data is obtained from Beth Israel Deaconess Medical Center. It was collected for 40,336 patients and posted on the Physio net Challenge 2019 and is available for the public.
The data consists of a of hourly vital sign summaries, laboratory values, and patients demographics. Dataset includes over 790215 data points. The data is extracted from the Electronic Medical Record (ERM) and underwent series of preprocessing steps prior to being published.
As mentioned the dataset has an hourly time sequence record for each patient. Each patient has up to 50 records spaced within one hour period. Each hourly record may or may not include measurements for all 41 features.

Classification problem could be approached in two ways. The time component could be taken as a part of the dataset, this provides us with additional information of how attributes of a patient change over time. Following this approach would require us to perform time series classification, using Long Short Term Memory Neural Networks. This presents a need for a lot of computing resources, since the number of data-points is significant. We can completely ignore the time component and treat each hourly record as its own entity. Following this approach the time dependency of features is lost. In this paper we'll focus on the second approach.

Dataset features are listed below. Note that the sepsis label is set to "positive" 6 hours in advance. This insures all of our predictions are made ahead of time, decreasing the probability of too late diagnosis.

$$t \geq t_s - 6 \longrightarrow SepsisLabel = 1 \quad (1)$$

$$t \ < t_s - 6 \longrightarrow SepsisLabel = 0 \quad (2)$$

Where t_s is time of sepsis diagnosis.

1) HR Heart rate [beats per minute]
2) O2Sat Pulse oximetry [%]
3) Temp Temperature [deg C]
4) SBP Systolic BP [mm Hg]
5) MAP Mean arterial pressure [mm Hg]
6) DBP Diastolic BP [mm Hg]
7) Resp Respiration rate [breaths per minute]
8) EtCO2 End tidal carbon dioxide [mm Hg]
9) BaseExcess Excess bicarbonate [mmol/L]
10) HCO3 Bicarbonate [mmol/L]
11) FiO2 Fraction of inspired oxygen [%]
12) pH [pH]
13) PaCO2 Partial pressure of carbon dioxide from arterial blood [mm H]
14) SaO2 Oxygen saturation from arterial blood [%]
15) AST Aspartate transaminase [IU/L]
16) BUN Blood urea nitrogen [mg/dL]
17) Alkalinephos Alkaline phosphatase [IU/L]
18) Calcium Calcium [mg/dL]
19) Chloride Chloride [mmol/L]
20) Creatinine Creatinine [mg/dL]
21) Bilirubin direct Direct bilirubin [mg/dL]
22) Glucose Serum glucose [mg/dL]
23) Lactate Lactic acid [mg/dL]
24) Magnesium Magnesium [mmol/dL]
25) Phosphate Phosphate [mg/dL]
26) Potassium Potassiam [mmol/L]
27) Bilirubin total Total bilirubin [mg/dL]
28) TroponinI Troponin I [ng/mL]
29) Hct Hematocrit [%]
30) Hgb Hemoglobin [g/dL]
31) PTT Partial thromboplastin time [seconds]
32) WBC Leukocyte count [count/L]
33) Fibrinogen Fibrinogen concentration [mg/dL]
34) Platelets Platelet count [count/mL]
35) Age [years]
36) Gender Female [0 / 1]
37) Unit1 Identifier for ICU unit [MICU] [0 / 1]
38) Unit2 Identifier for ICU unit [SICU] [0 / 1]
39) HospAdmTime Time between hospital and ICU admission [hours since admission]
40) ICULOS ICU length of stay [hours since ICU admission]
41) SepsisLabel [0 / 1]

C. Data Analysis

First we should address a common problem when dealing with real world medical data, that being the imbalance of datasets. Often the number of negative patients is much larger than the number of positive patients. This presents challenges when fitting our model and evaluating its performance. Models tend to learn a lot about the majority class and little about the minority class. While finding data distribution, we found that the dataset contains lots of missing values, this again is common with medical data. Due to the cost of particular measurements or their in-accessibility at a particular time, some measurements are missing. It is also worth noting that the patients are in hospital for different reasons, therefore some measurements may have a higher priority than others. Some measurements are also taken only in case patients state worsens. Percentage of missing values for each feature is shown in Fig. 1.

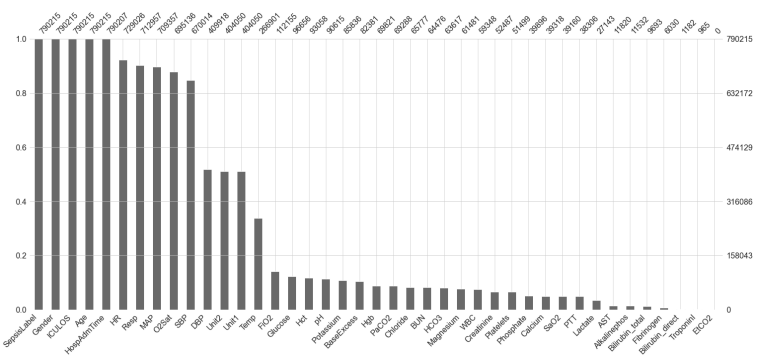

Fig. 1. Percentage of missing values

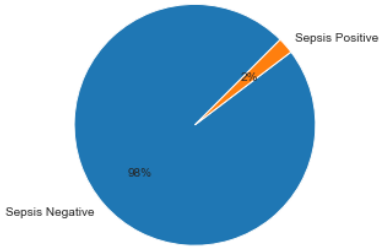

Fig. 2. Ratio between sepsis-positive and sepsis-negative patients

Looking at the distribution plots on Fig. 3. - 6. we notice that non-septic patient attributes tend to cover a larger spectrum of values. Here we point out a few features we decide to remove based on their statistical properties.

FiO2 is represented by somewhat discrete values with bi-modal distribution, this and the fact that only 20% patients have this attribute is enough of a reason to remove this feature. Further investigation shows this feature is strongly correlated with Bilirubin direct, which is missing for about 96% of patients, hence we keep only Bilirubin total.

There is a high correlation between DBP (diastolic BP), MAP (mean arterial pressure), and SBP (systolic BP). Mean arterial pressure is known to be primary indicator of patient state in near septic conditions, thus, both DBP and SBP may be disregarded as MAP is calculated from these two features.

Hemoglobin (Hgb) and hematocrit (Hct) values happened to be highly correlated. Since research articles about sepsis diagnosis mostly concentrate on hemoglobin out of these two, we decide to remove hematocrit. We also remove EtCO2, Unit1 and Unit2. To perform additional feature engineering such as selecting relevant features and finding relations between them, an expert in domain of sepsis diagnosis is required. Finally we take a look at a few interesting scatter plots. Scatter plots are

displayed on Fig. 8. - 10. Note that we ignore most of them, since the total number of possible scatter plots is given by the following formula.

$$m = \frac{n * (n-1)}{2} \tag{3}$$

where n represents number of features.

In our case m = 780. Here we select pairs of features that tend to separate the classes the best. We notice that among the best scatter plots, their does not exists a pair of features that completely separates the two classes.

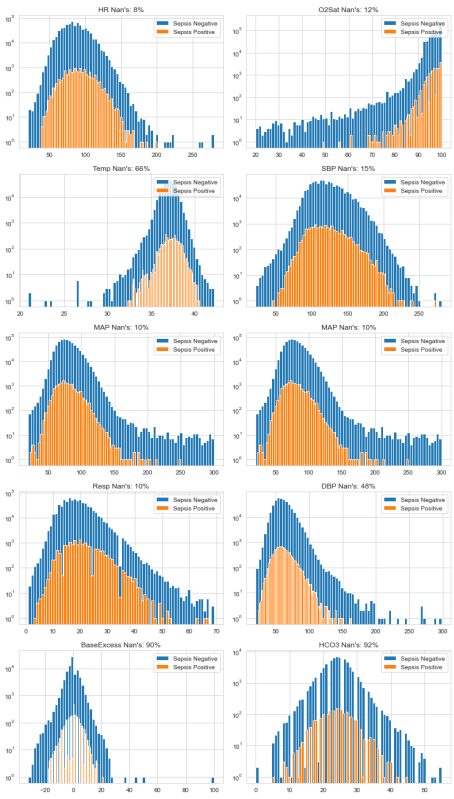

Fig. 3. Data distribution

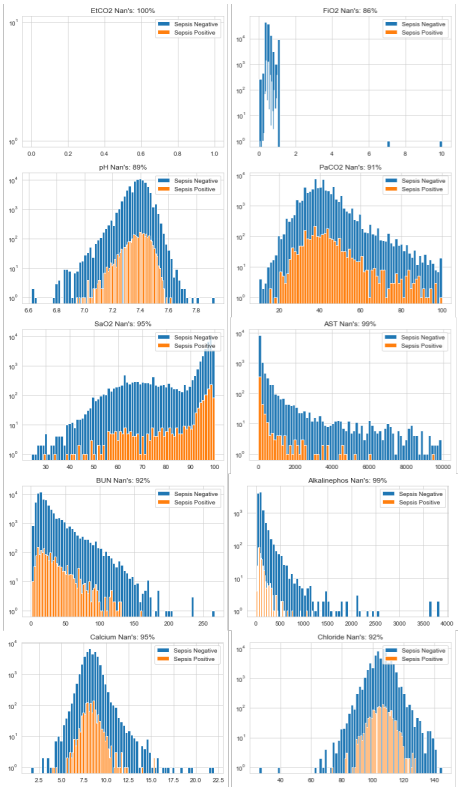

Fig. 4. Data distribution

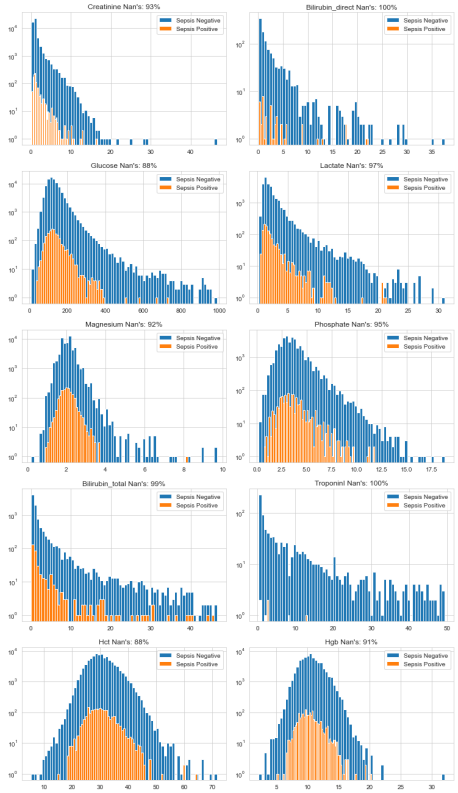

Fig. 5. Data distribution

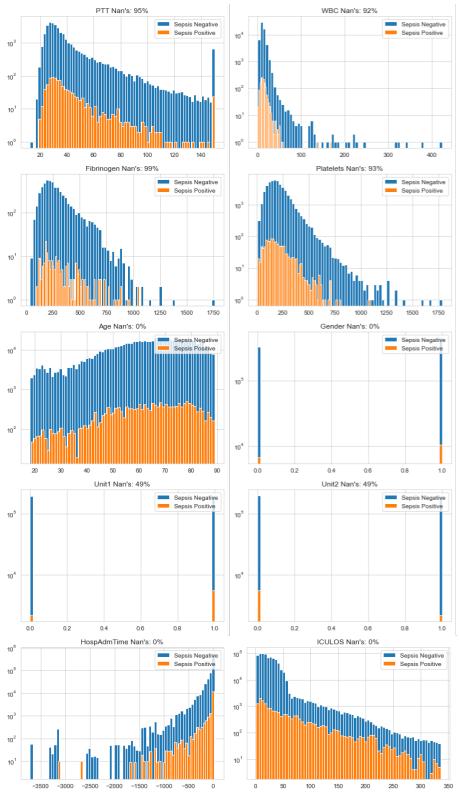

Fig. 6. Data distribution

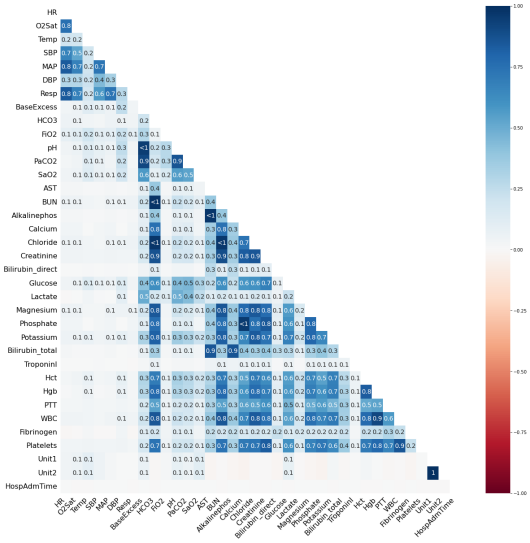

Fig. 7. Feature correlation map

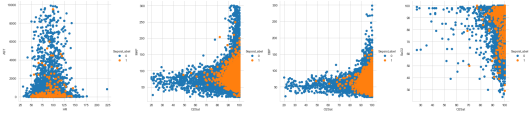

Fig. 8. Scatter plot

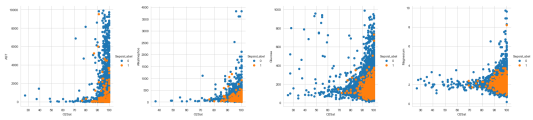

Fig. 9. Scatter plot

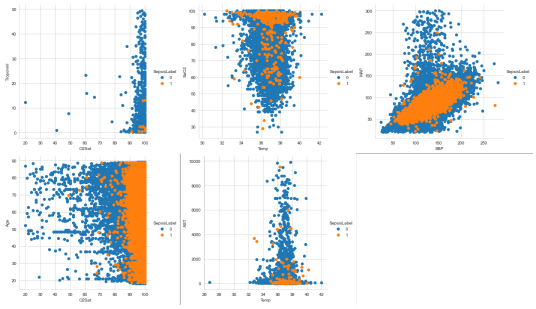

Fig. 10. Scatter plot

We should spend a few words on how sepsis is typically diagnosed and why its diagnosis is difficult. The following paragraph is taken from [2]. Infection is typically identified by three sets of information, listed bellow. However, not all mentioned symptoms are always present. E.g an immunosuppressed patient may not develop fever, and a source of infection is sometimes impossible to identify. Early on symptoms of sepsis can also be mistaken for influenza or other infections.

1) Clinical signs and symptoms of a host response: increase in white blood cell count, increase in concentrations of inflammatory markers.
2) The presence of signs of infection:
respiratory symptoms with abnormal chest auscultation and typical radiographic chest infiltrates, present signs of meningitis.
3) Proven microbiological invasion of a sterile environment.

D. Data Imputation

Performing data analysis we have tailored the number of features down to 30. Features we decide to keep are shown in Fig. 11. Since features with large percentage of missing values may still hold valuable information for classification we decide to impute missing data by choosing the closest value. This may aggravate the boarder between classes, but completely ignoring these features would very likely make it impossible to differentiate between patients.

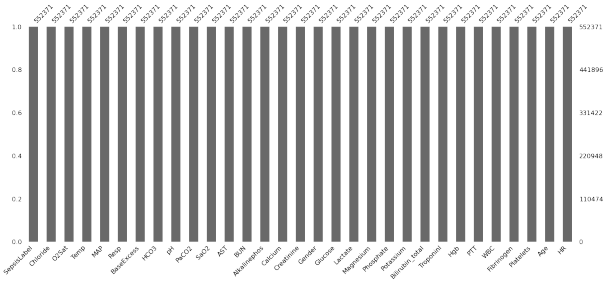

Fig. 11. Imputed features

E. Balancing the dataset

To balance the dataset we take advantage of two well known strategies: random-oversampling of minority class and Synthetic Minority Oversampling Technique (SMOTE) which produces new synthetic records. These synthetic records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class.

F. Model Evaluations

For the following analysis we treat each hourly record of a patient as a separate patient. We also remove features HospAdmTime and ICULOS indicating time.

1) Logistic Regression: We decide to start with Logistic regression, since it is easy to interpret and very efficient to train. Outputs have a probabilistic interpretation, and the algorithm can be regularized to avoid over-fitting.
Data is standardised (mean of features is brought to 0 and standard deviation is brought to 1). Results using random oversampling and SMOTE are displayed on Fig. 12. and Fig. 13.
In order to produces best models, a search for optimal hyper-parameters is performed. Hyper-parameters used include a list of regularization strength inverse values, and two penalty conditions (L1, L2).
Each combination of parameters is evaluated using 10 cross-validation folds.

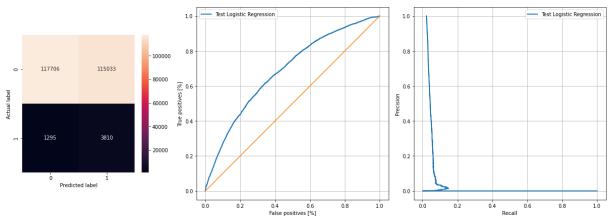

Fig. 12. Logistic regression random oversampling

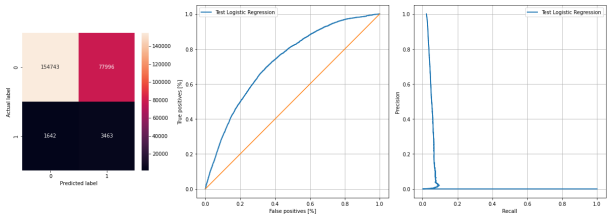

Fig. 13. Logistic regression SMOTE

We notice that SMOTE strategy produces significantly better results compared to over-sampling. It is interesting to see that SMOTE, performs better given the fact that our dataset is already synthetic to some degree. In general performance is bad, the ROC curve seems promising, but the area under the precision-recall curve is barely above no skill level. Looking at the confusion matrix we notice, the model has no trouble detecting sepsis negative patients, while detection of sepsis positive patients presents a challenge.

2) Random Forest: Random forest model can avoid the need for adding additional synthetic septic patients, by taking advantage of class weighting parameter of algorithm. We give a larger weight to the minority class. We train and evaluate two models, a baseline model and a model build using feature selection. Feature selection is performed to shrink the dimension of feature space. We keep features which provide the most information, this decreases the time needed to train the model and may also yield better results. Selected features are: HR, Temp, MAP, Resp, PaCO2, BUN, Chloride, Creatinine, Glucose, Potassium, Hgb, PTT, WBC, Platelets, Age. Results of both models are displayed in Fig. 14. and Fig. 15.

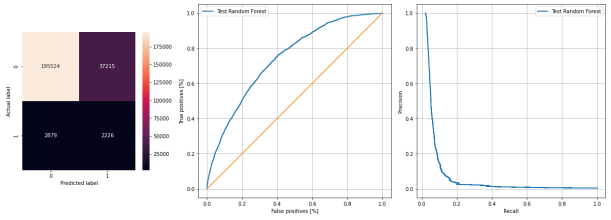

Fig. 14. Random forest baseline

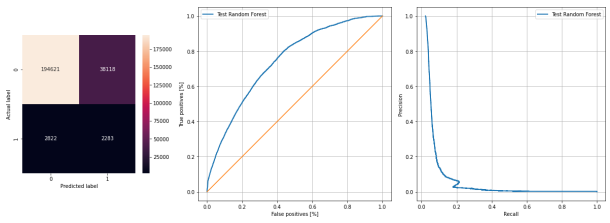

Fig. 15. Random forest feature selection

We find that Random Forest produces the largest AUC so far. Since the test dataset is imbalanced we should also pay attention to precision-recall curve[4]. Area under the precision-recall curve has increased quite a bit.

3) One Class Classification: One-class classification algorithms were originally designed for anomaly detection. In theory this strategy can be effective for imbalanced classification where there is no coherent structure to separate the classes that could be learned by a supervised algorithm. One Class models ignore the task of discrimination and instead focus on deviations from what is normal or what is expected

Assuming this may be the case we train and evaluate the Isolation Forest algorithm [3]. The algorithm is based on modeling the normal data in such a way to isolate anomalies that are both few in number and different in the feature space. The main benefit of the algorithm is its speed, when compared to one class classifiers based on support vector machines. Tree structures are created to isolate anomalies. The result is that, isolated examples have a relatively short depth in the trees, whereas normal data is less isolated and has a greater depth in the trees. One class classifiers offer a parameter to select the percentage of anomalies to consider, this offers direct control over the total anomalies the algorithm predicts. Behaviour of models can be altered to optimise a parameter relevant to specific application. A search for optimal hyper-parameters optimising f1 score (on septic patients) is performed using 10 cross validation folds. Optimal hyper-parameters increase the f1 score from 0.068 to 0.078. Confusion matrices of models are shown in Fig. 16. and Fig. 17.

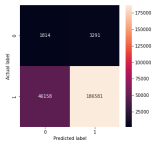

Fig. 16. Baseline Isolation Forest

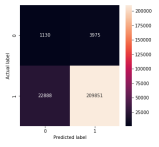

Fig. 17. Best Isolation Forest

Comparing the best septic f1-score to best f1-score of random forest we conclude that on this dataset, classic classification methods outperform the one class classifiers.

4) Neural Network: Based on current results it appears that our problem is highly non-linear. In this section we try multiple methods using Neural Network with two hidden layers. A dropout technique is used to prevent overfitting of the Neural network. A drop out probability of 50% is selected, therefor we decide to use a larger number of nodes. All methods below share the same neural network architecture. Data is first standardized, model is trained using training dataset and validation dataset. Metrics used to optimise during training include: TP, FP, TN, FN, Accuracy, Precision, Recall, ROC and PRC.

Layer (type)	Output Shape	Param #
dense_51 (Dense)	(None, 500)	15500
dropout_34 (Dropout)	(None, 500)	0
dense_52 (Dense)	(None, 250)	125250
dropout_35 (Dropout)	(None, 250)	0
dense_53 (Dense)	(None, 1)	251

Total params: 141,001
Trainable params: 141,001
Non-trainable params: 0

Fig. 18. Neural Network Architecture

5) Baseline Model: The baseline model is fitted using batch size of 60 and 100 epochs. Batch size appears to have a significant effect on classification results. Increasing the batch size increases the probability of containing a sepsis positive patient, on the other too large of a batch size will result in overfitting. Sizes were experimentally picked to decrease the level of overfitting. Baseline model results are shown in Fig. 19 and Fig. 20.

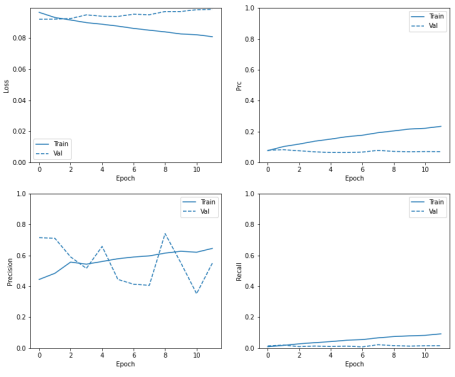

Fig. 19. Baseline model metrics

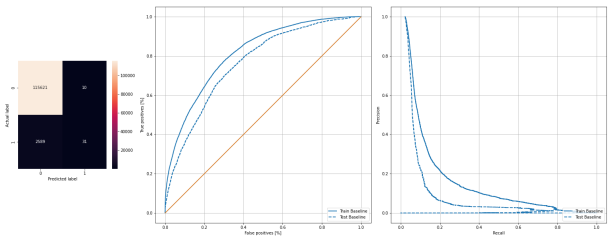

Fig. 20. Baseline model

6) Weighted model: To deal with imbalance we compute weights for both classes using the following formulas.

$$W_0 = \frac{1}{neg} * \frac{total}{2.0} = 0.51 \tag{4}$$

$$W_1 = \frac{1}{pos} * \frac{total}{2.0} = 22.96 \tag{5}$$

Where neg represents the number of sepsis negative patients

Where pos represents the number of sepsis positive patients

Results of weighted model are shown in Fig. 21. and Fig 22.

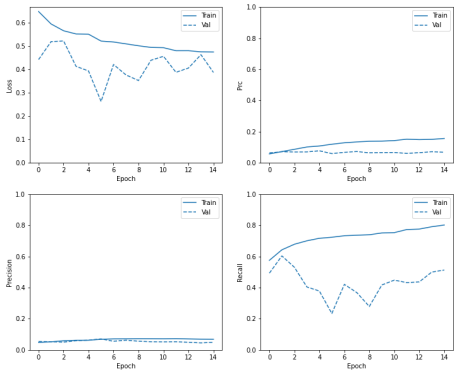

Fig. 21. Weighted model metrics

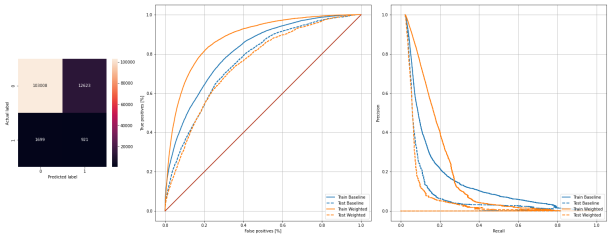

Fig. 22. Weighted model

We can see that the accuracy and precision of weighted model are lower because there are more false positives, but conversely the recall and AUC are higher because the model also found more true positives. Despite having lower accuracy, this model has higher recall (identifies more septic patients). Medical expert should consider the trade-offs between these different types of errors(recall and accuracy).

7) Class Weights with Feature Selection: In this section we try a simplified version of a method proposed in paper from references. We first scale our features, perform feature selection using Random Forest and Relief algorithm. Relief is an algorithm that takes a filter-method approach to feature selection that is notably sensitive to feature interactions. Features selected by the relief algorithm include: HR, Platelets, PTT, Glucose, HCO3, WBC, Hgb, Age, Resp, Potassium, MAP, BUN, Magnesium, Phosphate, BaseExcess. Batch-size of 60 and 100 epochs were used to train the models. Results using features selected by random forest are shown in Fig. 23. Results using relief to perform feature selection are shown in Fig. 24.

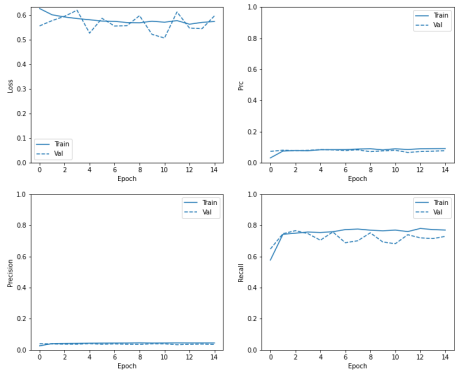

Fig. 23. Metrics of weighted model using features selected by random forest

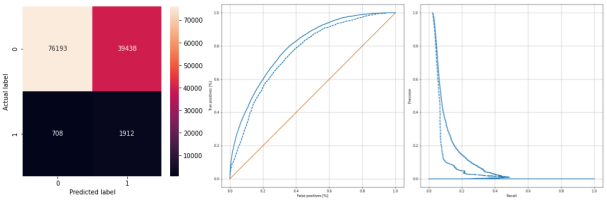

Fig. 24. Weighted model using features selected by random forest

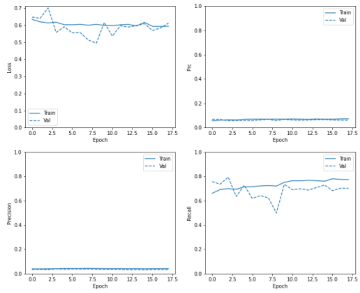

Fig. 25. Metrics of weighted model using features selected by relief

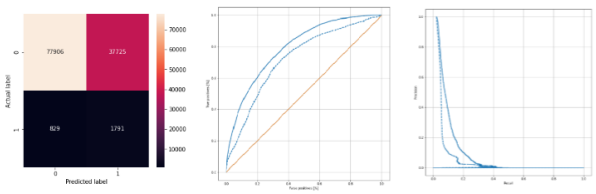

Fig. 26. weighted model using features selected by relief

G. Results

Following is a tabel of performance results. The meaning of the symbols used is as follows: LRO-logistic regression using oversampling technique to balance the data, LRS-logistic regression using Synthetic Minority Oversampling (SMOTE), RFB-random forest model using minority class weighting, RFF-random forest model using feature selection and minority class weighting, IF-Isolation forest classifier, NNB-baseline neural network, NNW-neural network using minority class weighting, NNF-neural network using features selected using random forest and minority class, weighting, NNR-neural network using features selected by relief algorithm and minority class weighting. Results of distinct models show a lot of variance in performance measurements. Weighting the minority class or using SMOTE brings down the accuracy and precision, because models detect more false positives, conversely the recall increases. Feature selection using both random forest and relief decreases the accuracy but increases the recall. As the complexity of a classifier increases the overfitting increases.

Model	Accuracy	Recall	Precision	F1	F2
r p LRO	0.51	0.75	0.03	0.06	0.14
LRS	0.67	0.68	0.04	0.08	0.17
RFB	0.83	0.43	0.06	0.10	0.19
RFF	0.83	0.45	0.06	0.10	0.19
IF	0.89	0.22	0.05	0.078	0.08
NNB	0.97	0.01	0.75	0.02	0.013
NNW	0.88	0.35	0.07	0.12	0.19
NNF	0.67	0.68	0.05	0.09	0.19
NNR	0.66	0.73	0.05	0.10	0.20

III. Conclusion

We have handled the missingness and imbalance of the dataset. Features with lots of missing values and little structure were removed. Strategies such as random forest feature selection and relief were used to select important features. We evaluated the following models: Logistic Regression on oversampled dataset and on SMOTE dataset, Random Forest using all features and only selected features, Isolation Forest and Neural Network. We aimed to predict the onset of sepsis by 6 hours, so far the models employed seemed to classify it partially. We conclude that classification on imbalanced datasets still present a challenge. Especially if the dataset also has a large percentage of missing values and therefore potentially synthetic data. Methods such as over-sampling and SMOTE do increase the recall, but the price we pay results in precision and accuracy drop. Effectiveness of SMOTE appears to be significant in case dataset is already synthetic too some degree. Oversampling the dataset tends to increase overfitting. Algorithms that provide the ability to weight minority classes perform competitively, they provide a benefit of fast training, which is an important role when dealing with large datasets. One class learning approach presents an alternative method, this approach may be more effective if the feature space of the classes where more distinct. Neural Network could be further tailored to insure as little overfitting as possible. All models could benefit of having a greater percentage of sepsis positive patients, especially given the fact that condition is so hard to diagnose. Next step of dealing with this dataset would be to go the time dependent route.

References

[1] Association, Information Košir, Domen Bosnic, Zoran Kononenko, Igor. (2013). The Use of Prediction Reliability Estimates on Imbalanced Datasets. 10.4018/978-1-4666-2455-9.ch035.
[2] Vincent JL. The Clinical Challenge of Sepsis Identification and Monitoring. PLoS Med. 2016;13(5):e1002022. Published 2016 May 17. doi:10.1371/journal.pmed.1002022
[3] F. T. Liu, K. M. Ting and Z. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.
[4] PLoS One. 2015; 10(3): e0118432. Published online 2015 Mar 4. doi: 10.1371/journal.pone.0118432
[5] https://www.physionet.org/content/challenge-2019/1.0.0/
[6] https://machinelearningmastery.com/one-class-classification-algorithmshttps
[7] https://www.tensorflow.org/tutorials/structured$_data/imbalanced_data$
[8] https://machinelearningmastery.com/how-to-control-the-speed-and-stability-of-training-neural-networks-with-gradient-descent-batch-size/
[9] https://machinelearningmastery.com/cost-sensitive-neural-network-for-imbalanced-classification/
[10] https://compsci682.github.io/notes/neural-networks-1/