

1 Using Machine Learning Algorithms to 2 predict sepsis and its stages in ICU patients

3 Nimrah Ghias¹, Shan Ul Haq², Huzifa Arshad¹, Haseeb Sultan¹, Farhan
4 Bashir¹, Syed Ameer Ghaznavi¹, Maria Shabbir³, Yasmin Badshah³, and
5 Mehak Rafiq¹

6 ¹School of Interdisciplinary Engineering and Sciences, National University of Sciences
7 and Technology, Pakistan

8 ²Xpertflow, National Science and Technology Park, Pakistan

9 ³Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and
10 Technology, Pakistan

11 Corresponding author:

12 Dr Mehak Rafiq¹

13 Email address: mehak@nust.rcms.edu.pk

14 ABSTRACT

15 Sepsis is blood poisoning disease that occurs when body shows dysregulated host response to an
16 infection and cause organ failure or tissue damage which may increase the mortality rate in ICU patients.
17 As it becomes major health problem , the hospital cost for treatment of sepsis is increasing every year.
18 Different methods have been developed to monitor sepsis electronically, but it is necessary to predict
19 sepsis as soon as possible before clinical reports or traditional methods, because delayed in treatment
20 can increase the risk of mortality with every single hour. For the early detection of sepsis, specifically
21 in ICU patients , different machine learning models i.e Linear learner, Multilayer perceptron neural
22 networks,Random Forest,lightgbm and Xgboost has trained on the data set proposed by Physio Net/
23 Computing in Cardiology Challenge in 2019. This study shows that Machine learning algorithms can
24 accurately predict sepsis at the admission time of patient in ICU by using six vitals signs extracted from
25 patient records over the age of 18 years. After comparative analysis of machine learning models , Xgboost
26 model achieved a highest accuracy of 0.98 , precision of 0.97, and recall 0.98 under the precision-recall
27 curve on the publicly available data. Early prediction of sepsis can help clinicians to implement supportive
28 treatments and reduce the mortality rate as well as healthcare expenses. **Keywords:** Sepsis , Prediction
29 , Machine learning, Comparative Analysis

30 INTRODUCTION

31 Sepsis is clinical syndrome caused by body's overwhelming which leads to tissue damage and organ
32 failure (9). In recent decades, sepsis still one of the life threatening disease in hospitals. It is most
33 commonly manifested by systemic bacterial infection that involves in the production of endotoxin, but it
34 can also be caused by fungal or viral etiology (4).It is linked with high morbidity rate, mortality rate and
35 responsible for hospital cost (12). Globally, an estimated 30 million people diagnosed sepsis in Intensive
36 Care Units and 6 million people died from sepsis each year (19). It is highly affected on adults and
37 children. The pathophysiological pathways of sepsis is very complex therefore it has variety of signs and
38 symptoms which are not easily detectable (15). The latest studies proposed that the mortality is increased
39 with every hour if the antibiotic treatment is delayed, because some patients having sepsis even at the
40 time of admission. Identifying risk factors earlier and commencing appropriate monitoring, before to any
41 clinical symptoms, would have a major influence on overall mortality and financial burden of sepsis (10).
42 Currently, available screening methods for sepsis i.e. systemic inflammatory response syndrome (SIRS),
43 modified early warning systems (MEWS), qSOFA etc are not enough for clear identification of sepsis
44 (8). Many researchers are concentrated on machine learning approaches for the excellent outcome and
45 high accuracy which is superior to the every disease severity scoring systems. Basically, machine learning

46 aims to develop algorithm that can learn and create models for prediction and data analysis which give
47 rapid outcomes (3). This current work was designed to adopt a real time machine learning algorithms
48 linear learner, xgboost, multilayer perceptron neural networks, lightgbm and random forest to detect
49 sepsis at the time when patient admitted in ICU, based on PhysioNet data collected from two hospitals. In
50 ICU, patients are admitted due to different reasons, the recognition of early sepsis with various disease
51 states (e.g inflammation) is quite challenging because every disease in ICU shows similar instances (e.g
52 dysregulated host response), clinical criteria (e.g change in vitals) and symptoms (e.g fever) (9). Machine
53 learning models have ability to learn predictive patterns in data that helps to handle the complexity and
54 wealth of digital patient data, which in turn give valid predictions about patient having sepsis. The
55 predictive patterns can be exposed either through supervised or unsupervised learning. The algorithms
56 that involve labeled training data (e.g. patients have sepsis or not) to predict outcomes for unforeseen data
57 is presented as supervised learning. In contrast, the data which has no labels and determine (known and
58 unknown) patterns in the data is included in unsupervised learning. Over the last years, many research
59 have used a range of computational models to deal with the difficulty in prediction of sepsis at its earlier
60 stage (9). The large number of features are retrieved from available attributes to train different machine
61 learning models and improve their performance. After verification of the proposed algorithms, through
62 5-fold cross validation method build the final ensemble model is applied on public challenge database and
63 make evaluation of this model on the hidden test set (17). The early detection of sepsis resulted in proper
64 monitoring and management of the patient leading to significant reduction in mortality rate (15).

65 **METHODOLOGY**

66 This research aims to predict sepsis at the time of patient's admission in ICU by applying machine learning
67 algorithms and extracted out the best model for the prediction. There are five steps involved to achieve the
68 goal.

- 69 1. Data Description
- 70 2. Tools Used
- 71 3. Data Preprocessing
- 72 4. Feature Selection
- 73 5. Machine Learning Algorithms

74 **Data Description**

75 The data is extracted from Physionet challenge 2019 which consist of 40336 PSV files, collected
76 from two different hospitals (Training set A which involved 20336 patients of hospital A and
77 Training set B involved 2000 patients of hospital B). Each file indicates hourly recorded data of
78 patients after admitting in ICU. The data includes 41 variables which consists of 26 laboratory values
79 (Measure of white blood counts, Bicarbonate, etc), eight vital signs (temperature, heart rate, oxygen
80 saturation, and systolic blood pressure etc), six demographics (gender,age, ICULOS, etc). The
81 last variable represents sepsis label 0 and 1. 1 means the sepsis has identified in patient on the
82 basis of sepsis 3 criteria. The data is highly imbalance that only 2932 out of 40336 patients has
83 sepsis. Additionally, there are many variables(26 out of 41) which have missing values more than
84 70 percent. For early sepsis prediction, the sepsis label has shifted forward for six hours in all data
85 (meaning that the label is set to 1 for six hours before it is officially identified).

86 **Tools Used**

87 There are many machine learning libraries i.e. scikit-learn, numpy, pandas, matplotlib which are
88 open source , and use for classification, clustering, regression and dimensionality reduction. Scikit-
89 learn is one of the most popular libraries which is used for evaluation of model and useful to extract
90 important features. If the dataset is highly imbalance then it is considered as quite challenging,
91 so to deal with the imbalance dataset there is library of Imbalanced-learn which offers multiple
92 resampling techniques i.e. SMOTE analysis.

93 **Data Preprocessing**

94 It is the most important phase in data formatting and data normalization. The review of data should
95 be carefully analyzed to avoid misleading results. Therefore, interpretation for accurate data should
96 be done before model building. The process of data preprocessing deals with redundant, missing and
97 noisy data and its strategies involved imputation of missing values and feature extraction. The large
98 number of missing values in the dataset was needed to be imputed for better prediction outcomes
99 by using different methods i.e. 0 imputation method, mean, median, mode and missforest method.
100 The main method to normalize the data is Min Max scalar or standard scalar and then Missforest
101 algorithm is considered as best imputation method as compared with other methods which has
102 validated by making histogram of every variable which shows normal distribution because better
103 imputation is the basic key for the better performance of model. Missforest can handle different
104 kind of data i.e. continuous and categorical. This algorithm doesn't require hyperparameter tuning.
105 It works on the basis of Random Forest which handle all missing value according to its requirement.
106 It predict the values on the basis of original data distribution and also useful to fix the imbalance
107 data (20).

108 **Feature Selection**

109 The mechanism of feature selection is used to filter out the most relatable features with the variable
110 which are needed to predict. The model accuracy can be effected by using inappropriate features
111 showing maximum outlier detection. This study has focused on six vital signs which are selected
112 on the basis of statistical analysis by using Z test having the idea that these vital signs are present in
113 all ICU patients and can be used for sepsis prediction. The correlation analysis has been used to
114 extract the features that were showing highly contribution as predicting variables.

115 **Machine Learning Algorithms**

116 There are many traditional methods i.e. laboratory test, qsofa score, SIRS etc. to detect sepsis but
117 delayed in detection due to unclear symptoms cause the high mortality rate and increase the cost of
118 hospitals therefore, there was need to predict sepsis earlier than clinical reports. For that purpose
119 different machine learning algorithms can be used for early detection with the high sensitivity and
120 specificity rate. For example, Xgboost, Random Forest and linear learner, LightGBM etc. Xgboost
121 is one of the best algorithm for the classification problem and shows accurate performance. It
122 shows iterative phenomena and combine all the results extracted from weak decision trees and
123 gives the best prediction. In every iteration it is focused on misclassified observations. It includes
124 the gradient boosted trees and construct the model. The prediction of sepsis was generated by
125 using six vital signs HR, temp, O2Sat, SBP and MAP at 6hr before prediction (1). Meanwhile,
126 XGBoost may process missing data automatically by assigning a default direction to null values.
127 To achieve the best XGBoost model performance, evaluation of hyperparameters was required ,
128 which included number of estimators, maximum depth and learning rates. The original dataset
129 was randomly partitioned into five subsets for this investigation. One-fold was utilized as a testing
130 subset, while the other four-fold were used to tune the hyperparameters, with 25 percent used for
131 calibration and the remaining 75 percent subjected to four-fold cross validation with grid search.
132 The hyperparameters selected that have the greatest area under the receiver operator characteristic
133 (18). Random forest was selected as the modern machine learning-based model, and it may be
134 viewed as an extension of existing tree based classifiers which is useful for classification and
135 regression problems. Random forest was chosen over other machine learning techniques (e.g
136 support vector machines) because it is similar to CART and has advantages when dealing with
137 EHR data. Random forest is an ensemble-based strategy that constructs several decision trees
138 (i.e., "forest") at the training data to offset the constraints of decision trees. Each tree is built from
139 a randomly selected subset of the original training data. A random subset of the entire number
140 of variables is evaluated at each splitting node. By adopting the mode of decision-making it can
141 reduce the problem of overfitting (14). LightGBM is great classifier for prediction which works
142 6 times faster than Xgboost. It learned about those attributes which having great contribution in
143 prediction (CHAMI et al.). It depends on histogram based algorithms which reduces consumption
144 of memory and speed up the training step. It combines advance communication networking for

145 parallel learning. That is why it is also known as parallel voting decision tree algorithm. In each
146 iteration, divide the training data into multiple machines and perform a local voting decision to
147 select the top-k attributes and a global voting decision to receive the top2k attributes(16).

148 Linear Learner algorithm is used for binary classification. It is having an option of normalization
149 for preprocessing. By turning on the normalization, it moves towards the smallest sample of
150 the data and find out mean value and standard deviation for every label and attribute .But for
151 binary classification, only features can be normalized. There are many optimization algorithms
152 are involved which can be used to take control for optimization processes and help to deal with
153 hyperparameters. When many models are trained in parallel manner, then they are compared with
154 validation set to check which model is optimal. The optimal model () gave the best F1 score ()
155 and accuracy () on the validation set. The other deep learning algorithm used for classification
156 in advance level is multilayer perceptron neural network which is also known as feed forward
157 neural network which involves input layer, hidden layer and output layer in which unlimited data
158 can be used. It doesn't only include vital signs, but also demographics or laboratory values. This
159 algorithm doesn't make any assumptions about distribution of data. The most attractive thing about
160 this technique is it can trained as numerical models on new data (6). It basically consists of nodes
161 or neurons having weights. Each neuron in MLP is connected to multiple of its neighbours, with
162 varied weights expressing the relative importance of the various neuron inputs on the other neurons
163 (7). The imbalanced number of neurons in hidden layer may cause the overfitting but there is no
164 specific method to find number of neurons. It is only dependent on trail and error method (11).

165 **EXPERIMENTAL RESULTS**

166 Correlation and Statistical analysis figured out those characteristics that have a significant impact on
167 sepsis prediction. The extracted attributes with more than 70 percent of missing values has dropped
168 and the rest of variables were used for the prediction which were imputed by using Missforest
169 algorithm.

170 Training set A include number of patients 790215 and Training Set B with number of patients () and
171 the third A/B combined dataset having patients 1552210 have used for model training by extracting
172 six vital signs. The groups include total (684107) number of males and (868103) number of females
173 and the age of patients which having sepsis is under 60-70 years. The summary of each dataset is
174 presented in Fig(). The data is standardized by performing SMOTE analysis give better outcomes
175 than without SMOTE. The performance of machine learning models are summarized on these three
176 datasets Each figure presents ROC curve of every model with the comparison of other datasets. The
177 graphs showing area under the precision recall curve and the area operating characteristic curve of
178 each dataset while table showing results of every model with Auc score, F1 score, precision and
179 recall.

180 **DISCUSSION**

181 Early sepsis prediction is significant problem but still challenging. This study proposed that machine
182 learning models shows high performance on prediction (AUROC 0.98) at the spot after patient's
183 data entry(10). Machine learning algorithms used hourly based data after patients admitted in
184 ICU to predict the prognosis of sepsis patients, the severity in condition of sepsis (i.e. septic
185 shock), and maximum length of stay of septic patients in ICU. Xgboost, Random forest, Lightgbm,
186 Linear Learner, Multilayer Perceptron neural networks classifiers had stronger predictive power,
187 with areas under the AUC score of 0.90, 0.92, 0.94 respectively. In early stage of sepsis, usage
188 of Random forest classifier allow to anticipate better ICU patients outcome, shows appropriate
189 medical measures and improve the treatment which improves prognosis (17). As many biological
190 events has happened in the pathophysiological of sepsis which leads to the disease processes and
191 health complications. (18). Its quite difficult to deal with disease complexity in ICU and imbalance
192 data, therefore, the advanced methods of machine learning presented the new scoring systems for
193 accurate prediction (13). The another interesting outcome is every model trained on combined
194 dataset Training set A and Training set B as well as on separate datasets and showing better results

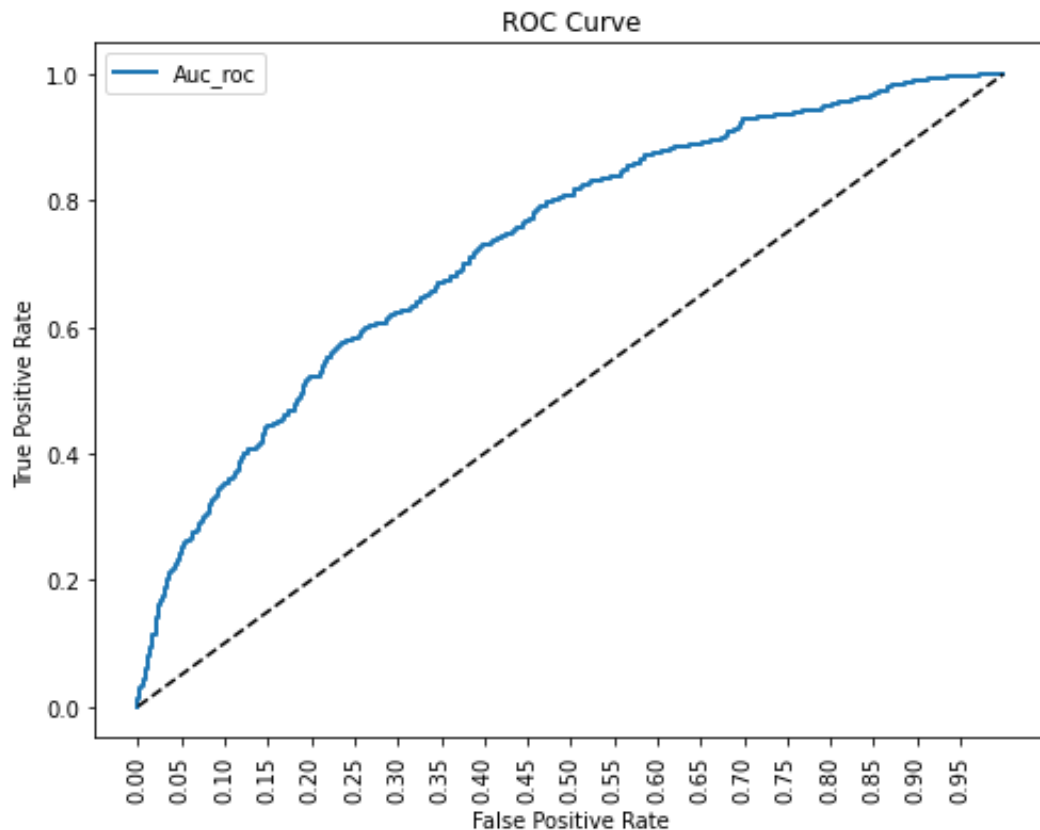


Figure 1. ROC curve

195 on training as well as on test dataset (5). Moreover, this study also shows the importance of each
196 feature that is having great impact on sepsis. The statistical analysis has been used for the purpose
197 of validation of each attribute on the basis of Z-test. The total number of septic and non septic
198 patients in dataset are examined and separate them in different classes and count the number of male
199 and female having sepsis and analyze the age which is more targeted due to sepsis. The prevalence
200 of sepsis is disproportionately higher in the elder patients and the age of a person is an independent
201 predictor of death. The elder patients are mostly non survivors of sepsis. This analysis is good
202 for better understanding about the data and helpful to know that sepsis mostly effects the female
203 as compared to male. The difference in male and female shows different hormone response to an
204 infection. The septic male and female have high estrogen level and shows the severity of illness in
205 females than males. Females with septic shock have high anti inflammatory mediators while males
206 have high tendency to maintain the health status. So by knowing the biological events it proved that
207 females have severe effect towards illness than male. After the statistical and correlation analysis
208 six vital signs has confirmed for the further process which are heart rate, temperature, oxygen
209 saturation, respiratory rate, mean arterial pressure, systolic blood pressure and diastolic pressure.
210 These variables having great impact in the prediction sepsis and can be used for model building.
211 This study shows the contribution in the comparison of different machine learning models and find
212 out the best model which can be deployed in hospitals. The model is trained on the features selected
213 from dataset. For the prediction of sepsis, every model has presented best performance by giving
214 ROC curve from (0.95 to 0.98). There is no limitation in distribution of features while using these
215 models therefore, they can used to tackle the large data as well. The evaluation of predictive model
216 occur by confusion matrix which compute the sensitivity, error rate, precision and specificity while

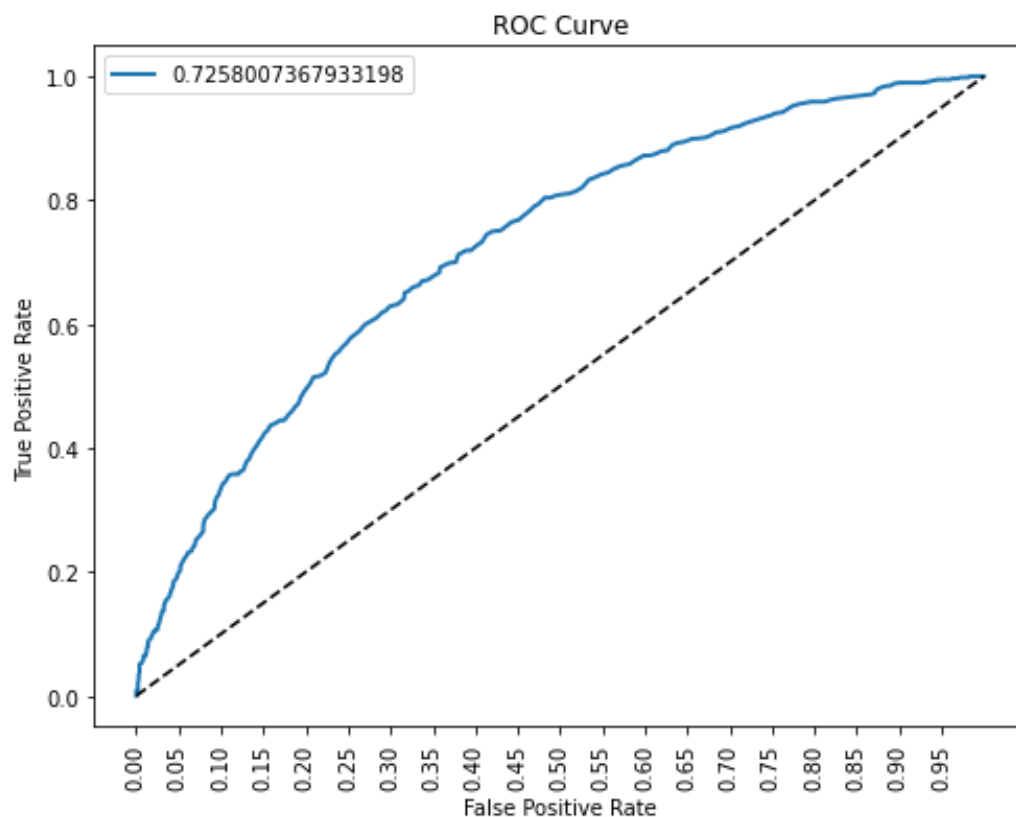


Figure 2. ROC curve

217 AUC is metric which differentiate the sepsis patients from other patients. In the comparison of
218 these ensemble models, Xgboost is more preferable than random forest because Xgboost shows
219 the integration of decision trees in sequential manner while random forest select each decision tree
220 individually and make a random subset for construction(?). This model could achieve highest ROC
221 curve because of better selection of features, dealing with imbalance data or overfitting through
222 smote analysis was the main key for the best prediction (5).

223 CONCLUSIONS

224 Sepsis is life threatening disease which cause of high mortality rate and morbidity in hospitals.
225 Early detection is a key to overcome the death rate, therefore this study showed the development
226 of fast and accurate machine learning algorithm for the prediction of sepsis which gives the better
227 results than the existing scoring systems. In addition, the comparative analysis has done between
228 five main models of machine learning by measuring their specificity and sensitivity. These models
229 has potential to use for commercial use in ICU's for sepsis prediction.

230 ACKNOWLEDGMENTS

231 So long and thanks for all the fish.

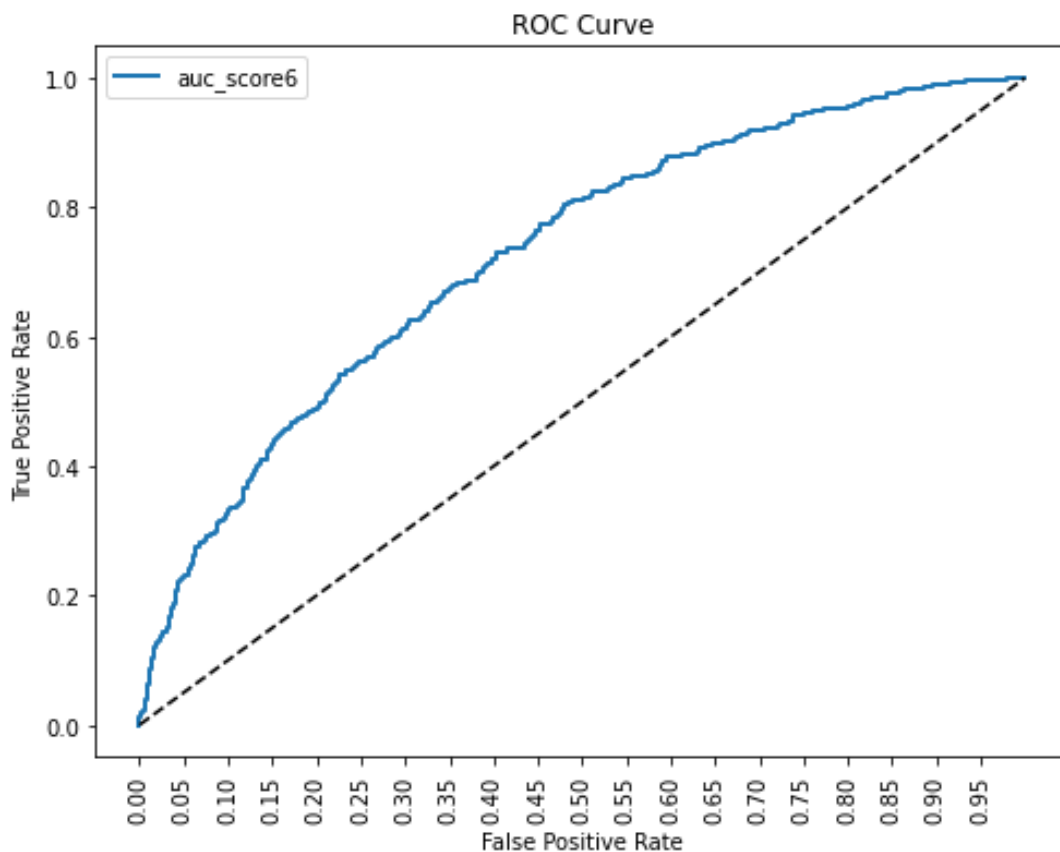


Figure 3. ROC curve

REFERENCES

232

- 233 [1] Burdick, H., Pino, E., Gabel-Comeau, D., Gu, C., Roberts, J., Le, S., Slote, J., Saber, N.,
234 Pellegrini, E., Green-Saxena, A., et al. (2020). Validation of a machine learning algorithm for
235 early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance
236 using a diverse dataset from 461 us hospitals. *BMC medical informatics and decision making*,
237 20(1):1–10.
- 238 [CHAMI et al.] CHAMI, S., Kaabouch, N., and Tavakolian, K. Comparative study of light-gbm and a
239 combination of survival analysis with deep learning for early detection of sepsis.
- 240 [3] Chibani, S. and Coudert, F.-X. (2020). Machine learning approaches for the prediction of
241 materials properties. *APL Materials*, 8(8):080701.
- 242 [4] Dolin, H. H., Papadimos, T. J., Chen, X., and Pan, Z. K. (2019). Characterization of pathogenic
243 sepsis etiologies and patient profiles: a novel approach to triage and treatment. *Microbiology
244 insights*, 12:1178636118825081.
- 245 [5] Fu, M., Yuan, J., Lu, M., Hong, P., and Zeng, M. (2019). An ensemble machine learning model
246 for the early detection of sepsis from clinical data. In *2019 Computing in Cardiology (CinC)*,
247 pages Page–1. IEEE.
- 248 [6] Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a
249 review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–
250 2636.
- 251 [7] Heidari, E., Sobati, M. A., and Movahedirad, S. (2016). Accurate prediction of nanofluid
252 viscosity using a multilayer perceptron artificial neural network (mlp-ann). *Chemometrics and*

927 150 355 232
0.720829200931985

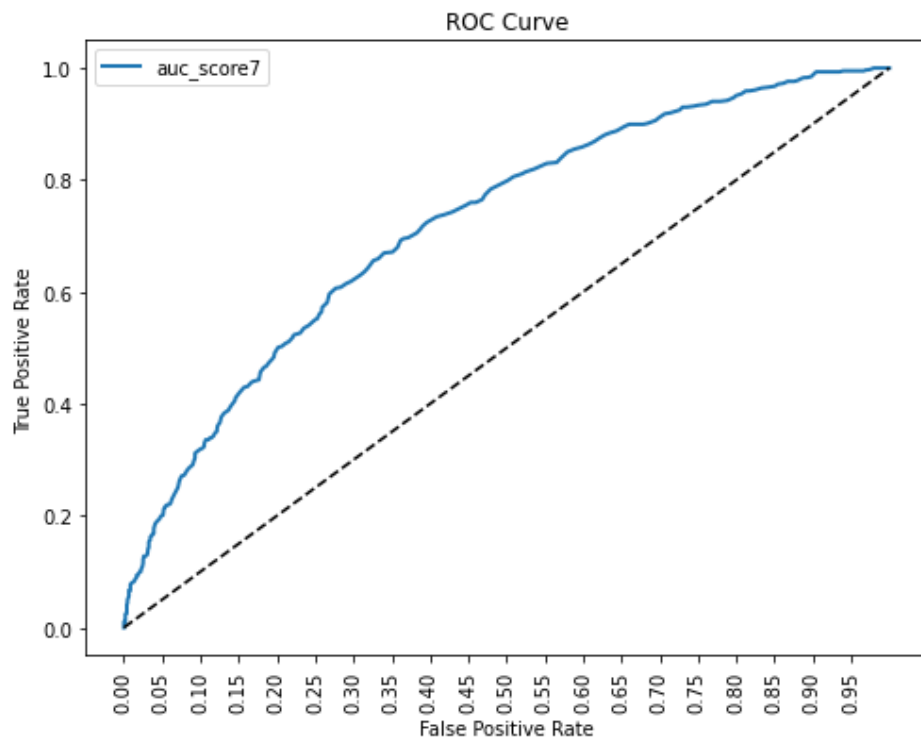


Figure 4. ROC curve

253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276

intelligent laboratory systems, 155:73–85.

- [8] Islam, M. M., Nasrin, T., Walther, B. A., Wu, C.-C., Yang, H.-C., and Li, Y.-C. (2019). Prediction of sepsis patients using machine learning approach: a meta-analysis. *Computer methods and programs in biomedicine*, 170:1–9.
- [9] Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., and Borgwardt, K. (2021). Early prediction of sepsis in the icu using machine learning: a systematic review. *Frontiers in medicine*, 8:348.
- [10] Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., and Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine*, 46(4):547.
- [11] Orhan, U., Hekim, M., and Ozer, M. (2011). Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10):13475–13481.
- [12] Parashar, A., Mohan, Y., and Rathee, N. (2021). Analysis of various health parameters for early and efficient prediction of sepsis. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, page 012002. IOP Publishing.
- [13] Su, L., Xu, Z., Chang, F., Ma, Y., Liu, S., Jiang, H., Wang, H., Li, D., Chen, H., Zhou, X., et al. (2021). Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models. *Frontiers in Medicine*, 8:883.
- [14] Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., and Hall, M. K. (2016). Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Academic emergency medicine*, 23(3):269–278.
- [15] Vincent, J.-L. (2016). The clinical challenge of sepsis identification and monitoring. *PLoS medicine*, 13(5):e1002022.

- 277 [16] Wang, D., Zhang, Y., and Zhao, Y. (2017). Lightgbm: an effective mirna classification method
278 in breast cancer patients. In *Proceedings of the 2017 International Conference on Computational*
279 *Biology and Bioinformatics*, pages 7–11.
- 280 [17] Yang, M., Wang, X., Gao, H., Li, Y., Liu, X., Li, J., and Liu, C. (2019). Early prediction of
281 sepsis using multi-feature fusion based xgboost learning and bayesian optimization. In *The IEEE*
282 *Conference on Computing in Cardiology (CinC)*, volume 46, pages 1–4.
- 283 [18] Yao, R.-q., Jin, X., Wang, G.-w., Yu, Y., Wu, G.-s., Zhu, Y.-b., Li, L., Li, Y.-x., Zhao, P.-y., Zhu,
284 S.-y., et al. (2020). A machine learning-based prediction of hospital mortality in patients with
285 postoperative sepsis. *Frontiers in Medicine*, 7:445.
- 286 [19] Zabihi, M., Kiranyaz, S., and Gabbouj, M. (2019). Sepsis prediction in intensive care unit
287 using ensemble of xgboost models. In *2019 Computing in Cardiology (CinC)*, pages Page–1.
288 IEEE.
- 289 [20] Zhao, X., Shen, W., and Wang, G. (2021). Early prediction of sepsis based on machine learning
290 algorithm. *Computational Intelligence and Neuroscience*, 2021.