Project report

Jun Chen, Shuai Liu

Abstract

Mortality prediction of hospitalized patients is an important problem recently. Over the past several decades, several scoring systems and machine learning mortality prediction models have been developed for predicting hospital mortality. By contrast Most ICU mortality prediction models in the literature have been created to provide real-time prediction on patients' mortality after ICU admission. Machine learning methods are prediction algorithms with potential advantages over conventional regression and scoring system. This study proposes a new machine learning approach XGBoost to address the work of predicting in-hospital mortality in the early stage of ICU stay using 1_day, 2_day and 3_day timeframe and to determine whether this model performs better than traditional prediction models. Data were extracted for the first 1_day, 2_days or 3_days since ICU admission for each ICU stay from MIMIC-III v1.4. The extracted features included physiological variables such as blood pressure, Glasgow coma scale, heart rate and demographic features such as ethnicity, gender. The data was split into two groups based on death or survival within timeframe and variables, eight predictive models including Random Forest, Logistic Regression, GradientBoost, AdaBoost, XGBoost, GaussianNB, KNN, MLP and XGBoost algorithm model were constructed by Python software. Then, the performances of the eight models were tested and compared by AUCs of the receiver operating characteristic curves, Accuracy score and MSE. The results show that the XGboost model performs best. For the mortality prediction, The AUC score of XGboost is 0.92, the accuracy is 0.92, the MSE is 0.28, for the death time multiclass classifier, the accuracy is 0.95, the MSE is 0.26. After that, we are using XGBoost to predict in-hospital mortality and death time in the three timeframes. The mortality prediction model trained on the 3_day ICU data has better performance, ROC is 0.94, whereas the death time multiclass classifier gives an effective base to provide an estimate of death time since ICU admission. micro-average ROC is 0.88 and macro-average ROC is 0.86. The conclusion is that using machine learning technique by XGboost, more significant prediction model can be built. This XGboost model may prove clinically useful and assist clinicians to better predict the Mortality.

Introduction

In the United States, each year over 30 million patients visit hospitals, 83% of which use an electronic health record (EHR) system [1]. Modern electronic healthcare records have an increasingly large amount of data, and the ability to automatically identify the factors that influence patient outcomes stand to greatly improve the efficiency and quality of care. This growth of digital clinical data gives a significant opportunity for data mining and machine learning researchers to solve pressing health care problems, such as risk assessment and early triage, identification of high-risk patients, prediction of physiologic decompensation and characterization of complex, multi-system diseases. Analyzing those huge data will help the medical diagnosis and treatment. In this project, we will use the Medical Information Mart for Intensive Care III database version 1.4 (MIMIC III v1.4) for the Study [2,3]. And we will predict patient's disease state and trajectory in ICU using advanced big data tools and novel machine learning techniques.

An intensive care unit (ICU) gives intensive treatment medicine for patients with life-threatening illness and injuries. Patients need to be treated and monitored to secure that they can be recovered. In a modern-day ICU, there are several types of devices such as pulse oximeter, ventilators, heart monitors, arterial lines, catheter, etc. that keep track of a patient's health records. In healthcare, predicting mortality in patients hospitalized in ICU is critical for evaluating severity of illness and judging the value of new treatments and health care policies. Over the past several

decades, several scoring systems and machine learning mortality prediction models have been developed for predicting hospital mortality. By contrast, early mortality prediction for intensive care unit patients remains an important challenge in current system. In this study, we will explore the MIMIC III database using various new machine learning models and techniques to predict early mortality prediction in ICU patients.

Early and accurate identification of patients with high risk of in-hospital death can help physicians in intensive care units (ICUs) make optimal clinical decisions. With the accumulation of big data and the development of techniques for data storage, machine learning methods have attracted considerable research attention [4–5]. Several innovative and practical machine learning methods such as random forest model [6], gradient boosting machine model [7] and the least absolute shrinkage and selection operator model [8] have been proposed, and these machine learning models have good prediction performance in medicine and other areas. Some machine learning-based models have been developed for ICU mortality prediction in the literature [9, 10]. In one research study, personalized mortality prediction was done by analyzing similar past patients. Several models were employed for the analysis. The average ROC values of the LASSO, RF, GBM were 0.829, 0.829, 0.845, respectively [11,12]. In another study, a neural network was used to predict mortality. Neural network resulted with an ROC score of 0.8445 and 0.86 for in-hospital mortality [13]. Some other studies have also assessed the advantage of nonparametric approaches. eXtreme Gradient Boosting (XGBoost) is a machine learning technique with the remarkable features of processing the missing data efficiently and flexibly and assembling weak prediction models to build an accurate one. As an open source package, XGBoost has been widely recognized in a number of machine learning and data mining challenges, for example, 17 solutions used XGBoost among the 29 challenge winning solutions published at Kaggle's blog in 2015 and the top-10 winning teams used XGBoost in KDD Cup 2015 [15]. For this big dataset, stability, speed and scale are features of the technology collection to be used, given the fact that software for health care systems should innated these. Another optimistic big data analytics tools for data science professionals that have appeared with those features are Hadoop and Spark. The power of Apache Hadoop and Spark is to perform data science tasks with stability, speed and accuracy. Another big superiority for Apache Hadoop and Spark is that it can combine ETL, data analytics, graph processing, machine learning and visualizations [14]. Using all those tools, computing resource and machine learning models are making the process and analysis for predicting early mortality in ICU much accuracy and faster.

Objectives

The goal of this study was twofold: firstly, we attempted to compare the performance of machine learning (XGBoost) model with traditional prediction models in the prediction of the 1_day mortality and death time in MIMIC-III. Eight models were trained on the extracted features of the study population for the specified timeframe. The model results were then compared and discussed in the study. Secondly, we planned to use the XGBoost to predict the mortality and death time in the early stage of ICU stay using 1_day ,2_day and 3_day timeframe.

Methods

Database Summary

We used the Medical Information Mart for Intensive Care III database version 1.4 (MIMIC III v1.4) for the study. MIMIC-III, a publicly available single-center critical care database which was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (BIDMC, Boston, MA, USA) and the Massachusetts Institute of Technology (MIT, Cambridge, MA, USA) [2-3]. MIMIC-III contains data associated with 46,520 distinct patients admitted to critical care units between 2001 and 2012. To form our study population, we have excluded ICU stays less than one hour to remove data due to unusual short stays and only consider adult patients with

age between 16 and 89. The final study population covered 49,632 ICU stays of 36,343 patients. The mean age of adult patients is 62.6 years, 57.8% patients are male, 70.4% is ethnicity white and in-hospital mortality is 11.6%. 82.3% is through emergency admission. The mean length of an ICU stay is 1.36 days. The database contains charted events such as demographics, vital signs, laboratory tests, fluid balance and vital status; documents International Classification of Diseases and Ninth Revision (ICD-9) codes; records hourly physiologic data from bedside monitors validated by ICU nurses; and stores written evaluations of radiologic films by specialists covering in the corresponding time period. After passing a training course on the website [2]. We were approved to extract data from this database for research purpose. The brief data summary is in the Table 1. The other summary graphs are in Appendix. Figure 1 shows the number of patients for in-hospital mortality by death time in days (1 to 60 days).

Table 1: Summary statistics of the study population

Variables	**Statistics** (mean ± std)
Age	62.61 ± 16.93
Gender	Male: 57.8% Female: 42.2%
Ethnicity	White: 70.4% Black: 7.1%
Admission Type	Emergency: 82.31%
Length of ICU stays	1.36 ± 1.06 days
In-hospital mortality ratio	11.6%

Figure 1. The number of patients for in-hospital mortality by death time in days (1 to 60 days).

Data Extraction

Our study is focus on to predict mortality and death time in the ICU stay. For each ICU stay, we have extracted data from the 1_day, 2_day and 3_day since ICU admission. We pull out the patients' features until they were discharged from the ICU. We then aggregated the feature values in the specified timeframe (1_day, 2_day and 3_day). There are altogether 123 extracted features, 6 categories, The Categories included Demographic and static features, Glasgow coma scale, Vital signs, Blood gases, Lab results, chemistry values and Urine output. In each Categories, there are also many physiological variables. For example, there are 5 variables in category demographic and static features. All these variables were aggregated and processed by minimum, maximum, and mean on the specified timeframe (1_day, 2_day or 3_day).

Model Evaluation

We propose to compare the performance of machine learning (XGboost) model with traditional prediction models in the prediction of the 1-day mortality in MIMIC-III. Secondly, we planned to use the XGBboost to predict the mortality and death time in the early stage of ICU stay using 1_day ,2_day and 3_day timeframe. If a patient is predicted dead in the first phase, the model would further predict an estimate of death time since ICU admission. Data were extracted for the first 1_day, 2_day or 3_day since ICU admission for each ICU stay from MIMIC-III database. Eight models were trained on the extracted features of the study population for the specified timeframe. The model results were then compared and discussed in the study. The dataset will be split into 80% training set and 20% test set. Hyperparameter tunings were done on 5-fold CV of the training set and the final evaluation of model performance was done on the test set. Multiple machine learning models in the study were evaluated and compared using the Area under the Receiver Operating Characteristic curve (ROC) on the test set. The ROC curve is the true positive rate against the false positive rate at various threshold settings. ROC provides a single measure of the diagnostic ability of a binary classifier. Other than the primary metric of AUROC, accuracy, RMSE were also used during the model testing stage to provide a full picture of model performance. And features importance from machine learning will be also reported.

In Phase 2, we label each data to one of the three classes specified. The distribution of data is following. Class 0 is death in one day. Class 1 is death between 1 to 7 days. Class 2 is death great than 7 days. XGBoost multiclass classifiers were then trained on the training set to predict the death time label using 1_day, 2_day and 3_day data respectively. Hyperparameter tuning was performed using grid search on 5-fold CV of the training set. The model performance of the best classifier resulted from the grid search under 1_day, 2_day and 3_day scenarios were then compared and evaluated on the test set.

Experimental Setup

Our study contained three stages of implementation: (1) Data summary and processing (**ETL process**) using Hadoop, Hive, Pig and PySpark on Local Docker environment (1 terabyte space, 16 GB RAM, and 8 processors, 6 GB GPU). (2) Modeling and Cross Validation (**Modeling**) on a local cluster (1 terabyte space, 16 GB RAM, and 8 processors, 6 GB GPUs). (3) Prediction mortality and death time using XGBboost on a local cluster (1 terabyte space, 16 GB RAM, and 8 processors, 6 GB GPUs). The unzipped dataset of MIMIC-III requires about 50 GB of space. Data output in the first stage is then used as the input for the model training in the second stage. We also used Python and packages such as Pandas and Scikit-learn for model testing, hyperparameter tuning and model evaluation.

Results.

Model comparisons

For Phase 1 model comparison results, we are using 1_day data as the practice dataset. Eight predictive models including Random Forest, Logistic Regression, GradientBoost, AdaBoost, XGBoost, GaussianNB, KNN, MLP and XGBoost algorithm model were constructed by Python software. showed good discriminatory power with AUCs of 0.91, 0.82, 0.907, 0.893, 0.919, 0.847, 0.664, 0.873. respectively (table 2). ROC curve graph was shown in the Figure 2. The XGBoost algorithm model showed the largest test AUC but the KNN model was the smallest. For the accuracy, XGBoost is highest with 0.923. MLP is the lowest with 0.861. for RMSE, XGBoost is the lowest with 0.278.

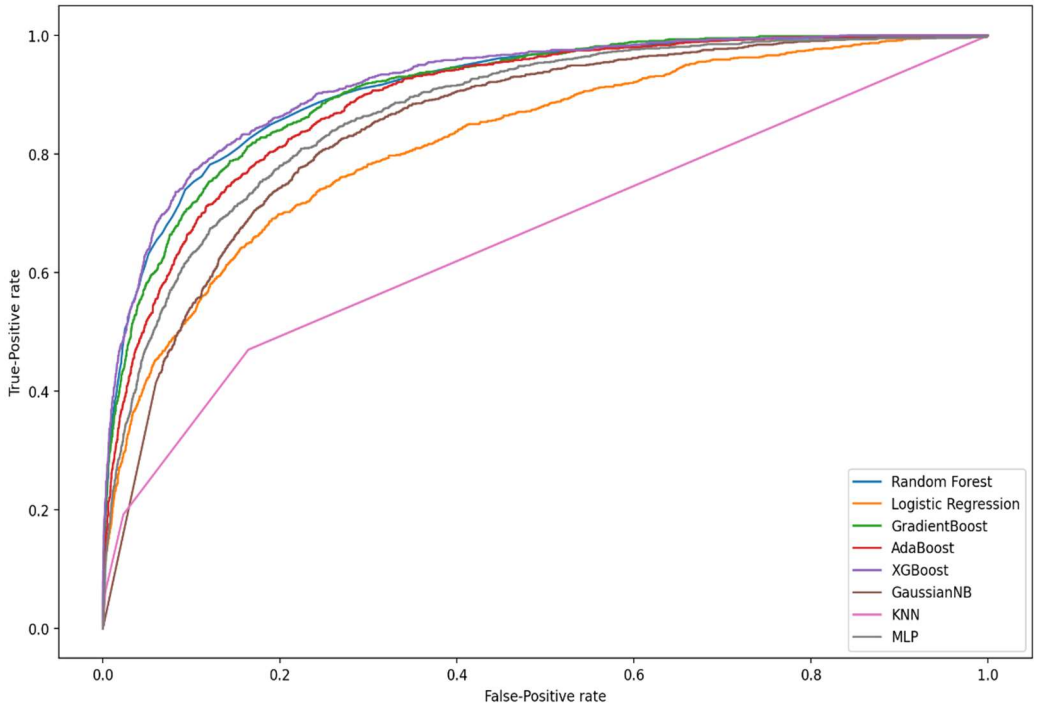

Figure 2. ROC curve graph for 8 models.

Table 2: Model performance of Roc_Auc_Score, Accuracy and RMSE for phase 1 data.

Methods	Roc_Auc_Score	Accuracy score	RMSE
Random Forest	0.9117	0.9168	0.2885
Logistic Regression	0.8212	0.9011	0.3145
GradientBoost	0.907	0.9165	0.289
AdaBoost	0.8928	0.9095	0.3008
XGBoost	0.9186	0.9226	0.2781
GaussianNB	0.8473	0.8228	0.4209
KNN	0.6638	0.889	0.3332
MLP	0.8733	0.8608	0.3731

For Phase 2 model comparison results, Table 3 compares the model performance of 8 models in Phase 2. showed good discriminatory power with accuracy score of 0.944, 0.939, 0.947, 0.945, 0.949, 0.807, 0.931, 0.942, respectively (table 3). The XGBoost algorithm model showed the highest accuracy score and lowest RMSE value. GaussianNB has lowest accuracy score and highest value for RMSE.

Table 3: Model performance of Accuracy and RMSE for phase 2 data.

Methods	Accuracy score	RMSE
Random Forest	0.9444	0.2835
Logistic Regression	0.9386	0.3092
GradientBoost	0.9469	0.2646
AdaBoost	0.9445	0.2801
XGBoost	0.949	0.2553

GaussianNB	0.8064	0.4959
KNN	0.9309	0.341
MLP	0.9416	0.2977

Model prediction

For Phase 1 model results, Figure 2 compares the model performance of the XGBoost classifiers separately trained using 1_day, 2_day and 3_day data in Phase 1 (Figure 3). After gridserach, Hyperparameter we are using is following (n_estimators=500, max_features=auto, criterion=gini, max_depth=30, bootstrap=True). The results show that data aggregation over wider timeframe gives slightly better result. Specifically, the XGBoost classifiers trained on 1_day, 2_day, 3_day data have AUROC 0.88, 0.90, 0.92 on the testset respectively. Figure 3 compares the ROC curves of the three classifiers and Appendix Figure 3 shows the feature importance of the top 20 features in Phase 1 data.

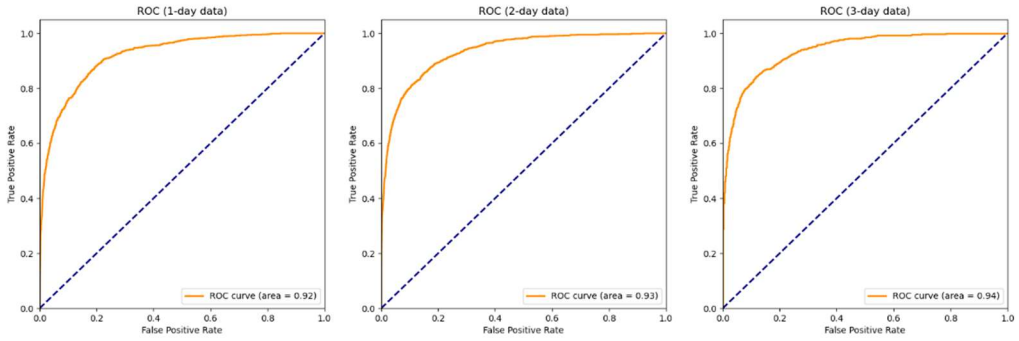

Figure 3. ROC curves of the three classifiers for Phase 1 data.

As for Phase 2 model results, Figure 3 compares the model performance of XGBoost multiclass classifiers separately trained using 1_day, 2_day and 3_day data in Phase 2 (Figure 4). The results show that data aggregation over 3_day timeframe gives slightly better result than those of 1_day and 2_day. Specifically, XGBoost classifier trained on 1_day, 2_day, 3_day ICU data have micro-average AUROC 0.77, 0.79, 0.82 on the test dataset respectively. Figure 4 compares the micro-averge, macro-average and ROC curves for individual classes. Appendix Figure 4 shows the feature importance of the top 20 features in Phase 2 data.

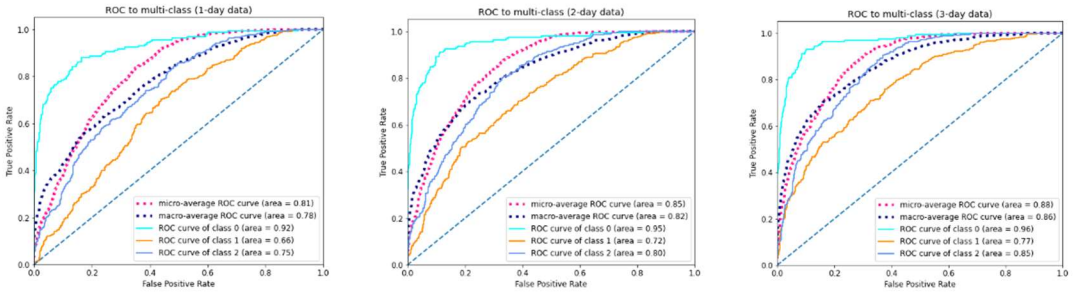

Figure 4. ROC curves of the three classifiers for Phase 2 data.

Discussion

In recent years, various machine learning algorithms, a subset of artificial intelligence and a data analysis technique that develops algorithms to predict outcomes by learning from data, have been investigated for early mortality and outperformed than conventional or classic statistic methods,

which could automatically analyze complex data and produce significant results. In this present study, the AUCs, accuracy score and MSRE we developed have demonstrated the benefit of using a XGboost model as opposed to the other machine learning for early prediction of mortality. Our results have showed that XGBoost has outperform the other methods based on all these three criterions. But there are several limits in XGBoost methods. For instance, the features selected were according to clinical experience but not algorithm; the representativeness of features may not clear in mortality and some important dynamic features were not included; besides, there were no validations for the XGboost model and no traditional regression analysis was used as a control.

The strength of this study was mainly that it was to predict the 1_day, 2_day and 3_day mortality of MIMIC-III patients in ICU and death time using the XGBoost mode. And compared to traditional regression analysis. We must admit some other limitations of our study and it may provide the potential improvement: firstly, because the data come from only one database and the most of patients were white, potential bias may occur; secondly, further exploration for the database was not performed, which may edge to the abandonment of some key features; Thirdly, the proposed model was not designed to be validated by developing dataset from the clinical data. Even so, we believe that the proposed model may contribute to further our understanding of the mortality prediction in ICU.

Conclusion

In conclusion, we have evaluated eight models to predict the mortality and death time categories. this study demonstrated that the machine learning based on XGboost algorithm does outperform conventional machine learning methods on both phases. This XGboost model may prove clinically useful and better predict the early mortality in ICU.

Video location: https://youtu.be/p9tSuyqjmdY

Code location: https://github.gatech.edu/jchen953/CSE6250_project

Challenges and learn: The big challenge for us is setting up the system and handle this big data. Another big challenge is grid search the best model in the machine learning models. It takes a very long time to run since this data is huge. What we learn from this project is that we have a quite good experiences for handling the big data using big data tools.

Timeline:

Task Description	Due date	Hours
Project Group Formation	2/28/2021	**3**
Project Proposal	3/15/2021	**15**
Data Processing	3/25/2021	**15**
Model Development	4/1/2021	**15**
Project Draft	4/15/2021	**15**
Project Final (final paper + code + presentation)	4/30/2021	**15**

Team contributions:

All the team members have contributed equally on this project.

References

1. Henry, J., Pylypchuk, Y., Talisha Searcy, M. & Patel, V. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015. ONC Data Brief 35, 2015.
2.Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.

3. Goldberger A, Amaral L, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet:components of a new research resource for complex physiologic signals. Circulation. 2000;101(23): e215–20.
4. Mahdi MA, Al_Janabi S. A novel software to improve healthcare base on predictive analytics and mobile services for cloud data centers. Cham: Springer International Publishing; 2020; 320–339.
5. Al-Janabi S, Mahdi MA. Evaluation prediction techniques to achievement an optimal biomedical analysis. Int J Grid Utility Comput. 2019;10(5):512–527.
6. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
7. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2000;29(5):1189–232.
8. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
9. Gurm HS, Kooiman J, LaLonde T, Grines C, Share D, Seth M. A random forest based risk model for reliable and accurate prediction of receipt of transfusion in patients undergoing percutaneous coronary intervention. PLoS One. 2014;9(5):e96385.
10. Van Poucke S, Zhang Z, Schmitz M, Vukicevic M, Laenen MV, Celi LA, De Deyne C. Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PLoS One. 2016;11(1):e0145791.
11. Kong G, Lin K, Hu Y Using machine learning methods to predict in-hospital mortality of sepsis patients in the icu. BMC Med Inform Decis Mak 2020; 251. 10.1186/s12911-020-01271-2
12. Fika, Sofia, et al. “A Novel Mortality Prediction Model for the Current Population in an Adult Intensive Care Unit.” Heart and Lung, 2018;47(1):10-15.
13. M. A. H. Zahid and J. Lee, "Mortality prediction with self normalizing neural networks in intensive care unit patients," 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, 2018; 226-229.
14. Harutyunyan, H., Khachatrian, H., Kale, D.C. et al. Multitask learning and benchmarking with clinical time series data. Sci Data,2019;6:96.
15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings
of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining-KDD 2016, San Francisco, CA, USA; 2016. p.785–94.

Appendix:

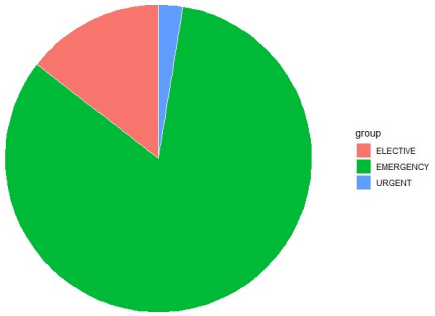

Figure 1. The distribution of patients by admission type.

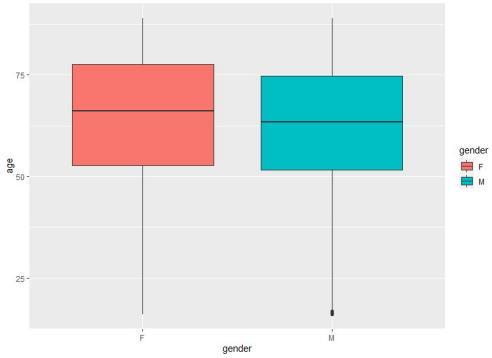

Figure 2. The distribution of patients age by gender.

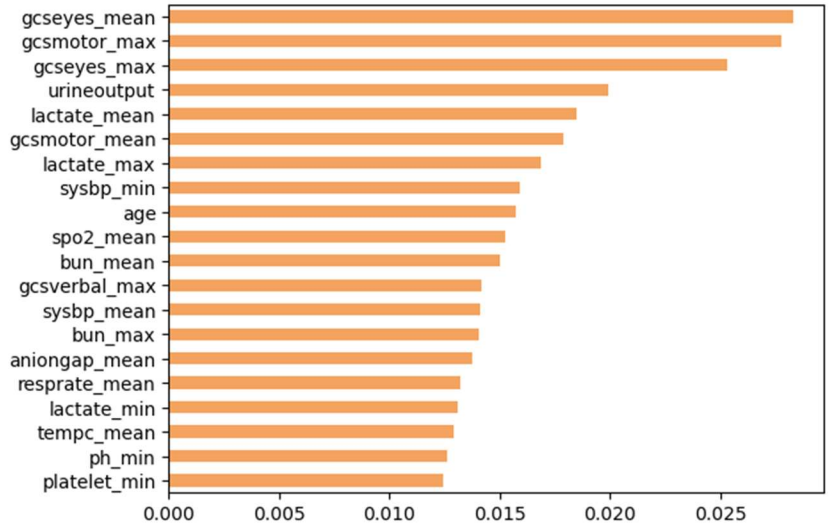

Figure 3. Top 20 features selected using XGBoost and the corresponding variable importance score for phase1 data. X-axis indicates the importance score which is the relative number of a variable that is used to distribute the data, Y-axis indicates the top 20 weighted variables.

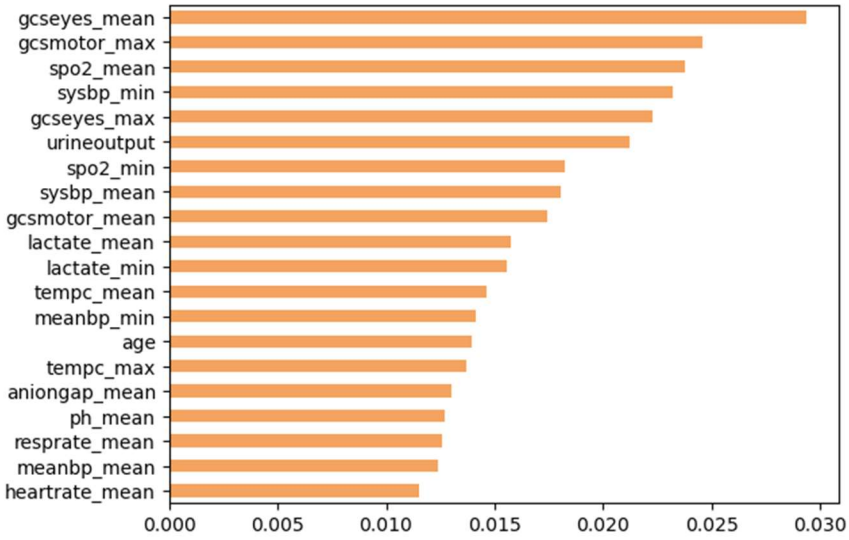

Figure 4. Top 20 features selected using XGBoost and the corresponding variable importance score for phase 2 data. X-axis indicates the importance score which is the relative number of a variable that is used to distribute the data, Y-axis indicates the top 20 weighted variables