Mortality prediction of sepsis

Mr.Gudala Karunakar[1], B.Jessy[2],B.Keerthi Sree [3], Ch.Shirisha[4], D.Reshma [5]

[2,3,4,5] UG Scholars, Department of CSE, ***MALLA REDDY ENGINEERING COLLEGE FOR WOMEN***, Hyderabad, Telangana, India.

[1]Assistant Professor, Department of CSE, ***MALLA REDDY ENGINEERING COLLEGE FOR WOMEN***, Hyderabad, Telangana, India.

Abstract

Sepsis is an important cause of mortality, especially in intensive care unit (ICU) patients. Developing novel methods to identify early mortality is critical for improving survival outcomes in sepsis patients. Using the MIMIC-III database, we integrated demographic data, physiological measurements and clinical notes. We built and applied several machine learning models to predict the risk of hospital mortality and 30-day mortality in sepsis patients. From the clinical notes, we generated clinically meaningful word representations and embeddings. Supervised learning classifiers and a deep learning architecture were used to construct prediction models. The configurations that utilized both structured and unstructured clinical features yielded competitive F-measure of 0.512. Our results showed that the approaches integrating both structured and unstructured clinical features can be effectively applied to assist clinicians in identifying the risk of mortality in sepsis patients upon admission to the ICU.

Introduction

Sepsis is a life-threatening organ dysfunction and a major public health issue. It is a common and economically important disease leading to 5.3 million death annually. The estimated overall mortality of sepsis patients is 30% [1-3]. Recently, sepsis was defined as a “life-threatening organ dysfunction caused by a dysregulated host response to infection” by The European Society of Intensive Care Medicine/Society of Critical Care Medicine Third International Consensus Definitions for Sepsis and Septic Shock task force (The Sepsis-3 task force) [2]. Early diagnosis and identification of sepsis are important to evaluate the patients’ status and improve their survival outcomes. Furthermore, because of the vague definitions of sepsis syndrome, unknown infection sources, and higher risk of mortality, developing an effective and reliable prognostic prediction model for sepsis patients is important. Such models could help to predict the prognosis of the patients more efficiently, inform the allocation of public health resources, and support clinical decision-making. Due to the increasing usage of electronic health records (EHRs), it is becoming easier to access comprehensive and extensive clinical data to predict health outcomes in the population. There are also many previous studies predicting mortality in patients [4-7]. However, many studies have largely focused on predicting mortality in specific population, such as elderly or patients who had cardiovascular surgery. These approaches could also draw meaningful conclusions in specific population, but it is also needed to build a mortality prediction model in general sepsis patients. Furthermore, many previous studies used several algorithm methods but did not include various laboratory information that is known to be effective for predicting the disease [8]. Finally, many of the mortality prediction studies have been conducted based solely on structured EHR data and did not incorporate unstructured clinical notes which could be ubiquitously utilized in other medical institutions as well [9]. Therefore, in this study we included a wide array of predictors including demographic characteristics and various physiological factors from the laboratory test that are recorded in the EHR, to predict mortality in sepsis-3 patients. Moreover, we used structural data (e.g., physiological variables) as well as unstructured intensive care unit (ICU) clinical notes data to construct the prediction model. These clinical notes are written by many clinical experts, including physicians and nurses, and could provide a comprehensive picture of patients’ pathological statuses and aid in the development of a powerful model to predict the mortality in sepsis patients.

Literature Review

A diagnosis system based on artificial intelligence (AI) is shown to be effective in many medical fields. In the area of diagnosis, prognosis, and treatment of sepsis, machine learning algorithms used include supervised learning and reinforcement learning [2–5]. For example, Beck et al. [6] develop the C-Path (Computational Pathologist) system to automatically diagnose breast cancer and predict whether patients will survive or not by examining breast tissue imaging.

The main two challenges in the current research include the use of different physiological indicators and modeling efficient machine learning algorithms for the diagnosis, prognosis, and treatment of sepsis. Similarly, in order to predict sepsis in advance, it is also crucial to choose appropriate variables and design valuable algorithms in the clinical setting.

The input variables of the model are physiological indicators and the output variable is whether the patient would suffer from sepsis several hours later. Specifically, the input variables generally include vital signs like heart rate, oxygen saturation, and body temperature; biomarkers like procalcitonin and interleukin-6; laboratory values like bicarbonate and creatinine; and demographic variables like sex and age. In most cases, the variables include lots of missing values, such as that in MIMIC III (Intensive Care Medical Information Market Database), which has been used in many studies. Among most researches, variables with lots of missing values are excluded from predictors, so valuable information may be lost as a result. Several studies use imputation and mean filling methods to fill in missing values, but this may also lead to selection bias or mixtures of confounding factors. The data preprocessing method needs to be considered according to the characteristics of different data sets.

Common ways to deal with missing values are missForest [7], KNNimpute [8, 9], and so on. Other ways are also proposed. For instance, Desautels et al. [10] proposed the InSight algorithm by using easy-to-monitor patient vital signs data and an integrated tree boost algorithm to train the model so as to simplify the types of input variables as much as possible. The final simplified input variables include vital signs (systolic blood pressure, pulse pressure, heart rate, respiratory rate, temperature, and peripheral capillary oxygen saturation (SpO2)), patient age, and Glasgow coma score (GCS). Its AUC indicating discriminative power between infected and noninfected patients reaches 0.880. Taneja et al. [11] make a detailed comparison of input variables such as vital signs and biomarkers and predict sepsis risks 4 hours in advance. The vital signs and biomarkers are separately used as input variables to train the model to obtain the AUC score, and then, they are both used as input variables to train the model to compare the effects. The final feature importance is listed in order as vital biomarkers and vital signs.

The machine learning algorithms generally include support vector machines, gradient boosting trees, random forests, Lasso regression, and neural networks. Among them, support vector machines and gradient boosting trees have shown good performance. The model with better prediction ability will be further tested and improved for clinical service so that clinicians can make better decisions in sepsis early diagnosis. Taneja et al. [11]

compared the predictive abilities of five machine learning models, including logistic regression, support vector machines, random forests, Adaboost, and Naive Bayes. Among them, the support vector machine algorithm and Adaboost algorithm have the highest AUC scores. The other models in use also include deep learning methods and biological methods. For example, Scherpf et al. [12] used a recurrent neural network (RNN) to conduct experiments on the sepsis data set provided by the MIMIC III platform. Nemati et al. [13] used a proportional hazard model to predict sepsis several hours in advance. Lin et al. [14] used the convolutional LSTM model, the random forest model selected by Lamping et al. [15], and the Gaussian process-based RNN model used by Hariharan [16].

The above studies have shown good performance in the field of sepsis prediction. However, the amount of data used in these researches are shrunk, as most of the missing values are processed by direct deletion or forward filling, and the explanatory ability of the model is also limited. It is challenging to transfer these methods into clinical practice for the following detailed reasons. (1) A unified data set is lacked. Researchers use data from different patient groups, for example, the MIMIC public database or other independent hospital data sources. The clinical variables they select to generate models differ and the scale of data differs a lot as well. (2) The premise and indicators of prediction settings vary, such as clinical standards for sepsis, observation windows, and evaluation indicators.

Above all, it is still not possible to do full validations for sepsis prediction in different groups with current machine learning methods and evaluate their generalizability. In addition, many of the machine learning models are complex and hard to be explained. Clinicians lack tools to interpret this “black box” model in clinical practice. This study is committed to digging out the most effective information from large-scale data. In terms of the interpretability ability, a metric called SHAP value is used in this study which can help models break the “black box” barriers and have good interpretability.

Specifically, this research develops machine learning models with good generalization ability and clinical interpretability by generating two data preprocessing methods based on XGBoost and LightGBM algorithms, which can be used to predict early sepsis 6 hours in advance, to assist clinicians in early diagnosis, intervention, and treatment. (1) In the mean processing method, it is explored whether or not the model predictive ability will be improved by extracting mean vectors. After dividing the early warning period into 2 hours or 3 hours window, it is discussed about the relationship between the extent of category imbalance and the model’s predictive ability. (2) In the feature generating model, the prediction performance of raw variables trained in different models are compared with those extra with different types of newly generated features in the relationship between model performance and model complexity.

The rest of the research is arranged as follows. In Section 3, materials and methods are given. The data used for prediction are introduced, followed by the two data processing methods and the prediction process. Section 4 reports the results of predictive analysis and explores the complexity of data preprocessing, as well as the number and types of new features generated which affected the model’s prediction ability. Section 5 gives the conclusion and future work.

System Analysis:

Exisiting System

Sepsis is a very dangerous and serious condition caused by an infection that leads to tissue damage and organ failure due to an increase of chemicals in the bloodstream. It is crucial to provide fast and efficient treatment for the patient. Currently in clinical practice there are used different scoring systems to diagnose sepsis and the most important score system today is the sequential organ failure assessment score also abbreviated to SOFA score. It is based on various physiological and laboratory measures which are taken from blood samples and by analyzing if they are within a normal range or not.

Proposed System

The mechanism of feature selection is used to filter out the most relatable features with the variable which are needed to predict. The model accuracy can be effected by using inappropriate features showing maximum outlier detection. This study has focused on six vital signs which are selected on the basis of statistical analysis by using Z test having the idea that these vital signs are present in all ICU patients and can be used for sepsis prediction. The correlation analysis has been used to extract the features that were showing highly contribution as predicting variables.This study shows the contribution in the comparison of different machine learning models and find out the best model which can be deployed in hospitals. The model is trained on the features selected from dataset. For the prediction of sepsis, every model has presented best performance by giving ROC curve from (0.95 to 0.98). There is no limitation in distribution of features while using these models therefore, they can used to tackle the large data as well.The evaluation of predictive model occur by confusion matrix which compute the senstivity, error rate, precision and specificity while AUC is metric which differentiate the sepsis patients from other patients.

Results :

Mortality Prediction of Sepsis

In this project we are predicting mortality rate for sepsis patients admitted under ICU. For prediction we are employing machine learning algorithms such as Random Forest and XGBOOST and to train both algorithm we are using MIMIC3 dataset. To evaluate performance of both algorithms we are training them on 80% dataset and then testing their prediction accuracy on 20% test data. To evaluate performance we have used other metrics called Confusion Matrix, ROC graph, Precision, recall and FCSORE.

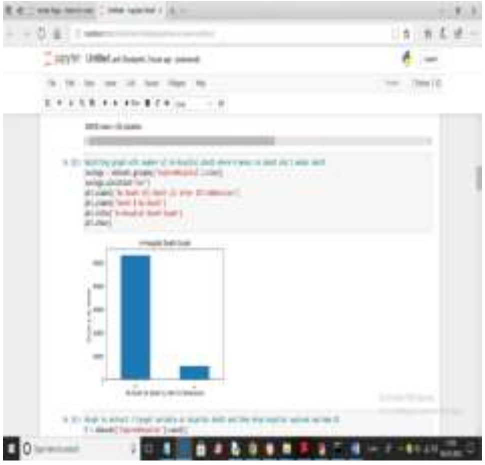
Below screen showing dataset details used to train above algorithms

In above dataset screen first row contains column names and remaining rows contains dataset values and from above dataset we are using Hospital Expired column as Target Value.

We have coded this project using JUPYTER notebook and below are the code and output screens with blue colour comments

In above screen importing all python packages

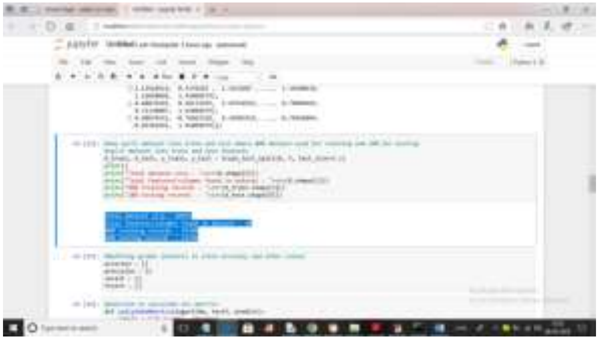
In above screen reading and displaying dataset values



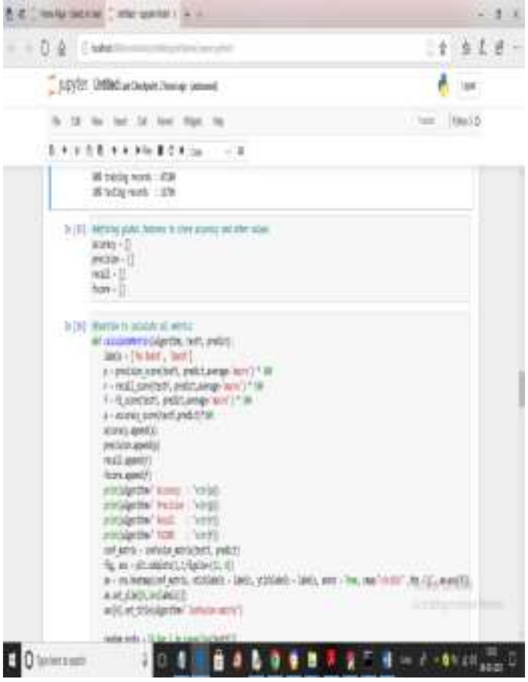
In above screen finding and plotting graph of ‘No-Death & in-hospital death’ where 0 means no death and 1 means death which is in X-axis and death count in y-axis

Dataset contains both numeric and non-numeric data but ML algorithms take only numeric data so by applying Label encoding class we are converting all non-numeric data to numeric data which you can see in above screen and after applying Label Encoding we can see all values are converted to numeric format

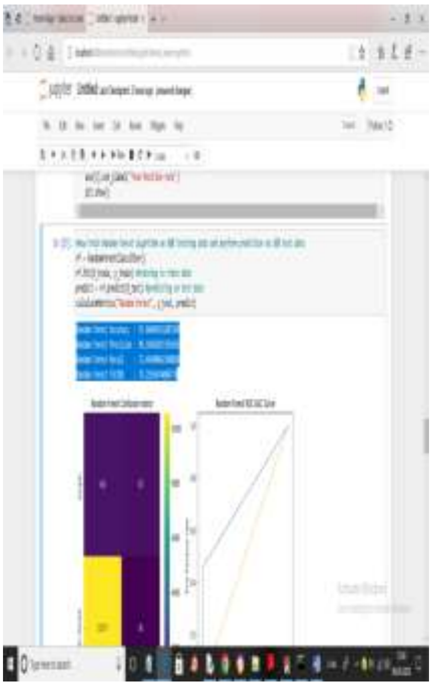
In above screen we are processing dataset by using data shuffling and normalization technique and then displaying normalized values



In above screen we are splitting dataset into train and test and in blue colour text you can see output of training and testing records

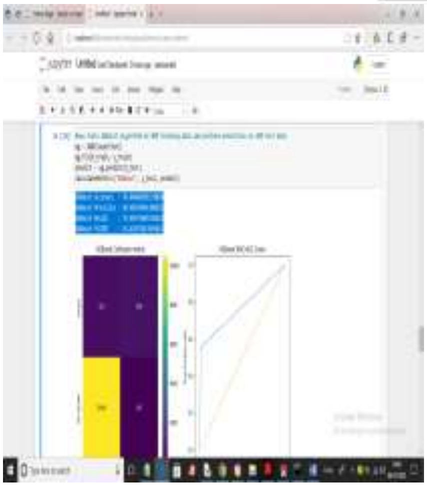


In above screen defining function to calculate accuracy, precision, confusion matrix and other metrics

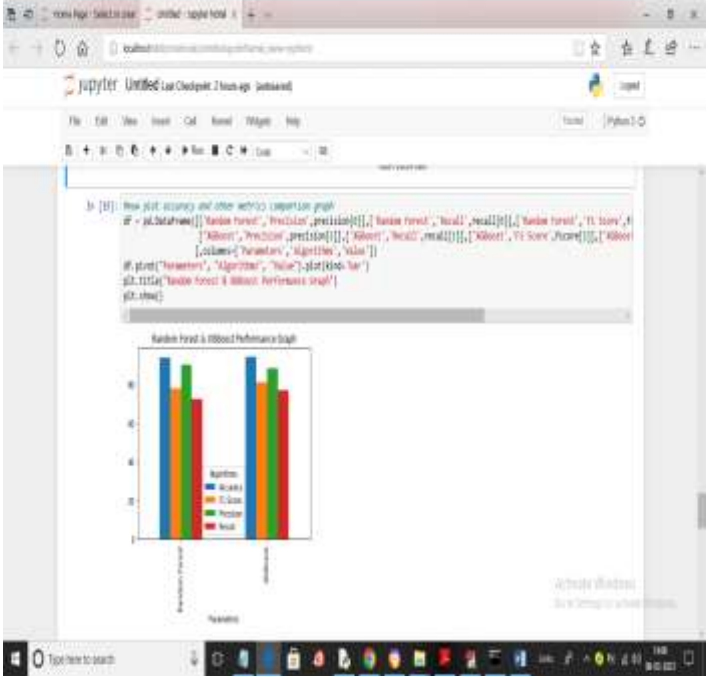


In above screen we are training Random forest algorithm on training data and performing prediction on test and we got its accuracy as 93% and below are the graph

In above confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and yellow and blue boxes on its diagnol contains correct prediction count and both blue boxes in diagnol contains incorrect prediction count. In ROC graph x-axis represents False Positive Rate and y-axis represents True Positive Rate and if blue line comes on top of orange line then all predictions are true and if comes below orange line then all predictions are wrong

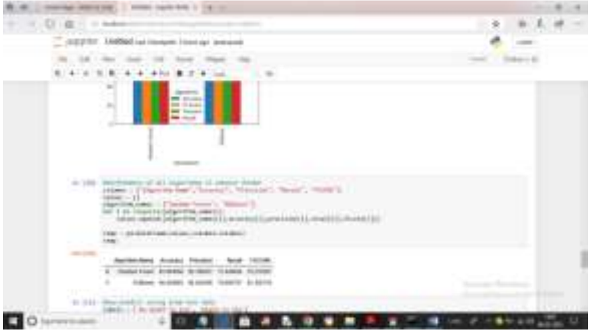


In above screen we are training XGBOOST and we got its accuracyas 94% and we can see graphs also



```
In [18]: #now plot accuracy and other metrics comparison graph
         df = pd.DataFrame([['Random Forest','Precision',precision[0]],['Random Forest','Recall',recall[0]],['Random Forest','F1 Score',f
                            ['XGBoost','Precision',precision[1]],['XGBoost','Recall',recall[1]],['XGBoost','F1 Score',fscore[1]],['XGBoost
                           ],columns=['Parameters','Algorithms','Value'])
         df.pivot("Parameters", "Algorithms", "Value").plot(kind='bar')
         plt.title("Random Forest & XGBoost Performance Graph")
         plt.show()
```

In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in both algorithm XGBoost got high accuracy



```
   Algorithm Name  Accuracy  Precision  Recall  FSCORE
0   Random Forest  [illegible]
1         XGBoost  [illegible]
```

In above screen we can see both algorithms performance in tabular format

In above screen we are reading test data and then predicting death or no death and in above screen in square bracket we can see test values and after arrow symbol =➔ we can see predicted values as 'Death in ICU' or 'No Death in ICU'

Conclusion:

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

Refrences:

1 K. E. Rudd, S. C. Johnson, K. M. Agesa et al., "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study," The Lancet, vol. 395, no. 10219, pp. 200–211, 2020.
View at: Publisher Site | Google Scholar

2 L. Su, Z. Xu, F. Chang et al., "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, vol. 8, 883 pages, 2021.
View at: Publisher Site | Google Scholar

3 K. C. Yuan, L. W. Tsai, K. H. Lee et al., "The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit," International Journal of Medical Informatics, vol. 141, Article ID 104176, 2020.
View at: Publisher Site | Google Scholar

4 J. E. García-Gallo, N. J. Fonseca-Ruiz, L. A. Celi, and J. F. Duitama-Muñoz, "A machine learning-based model for 1 year mortality prediction in patients admitted to an intensive care unit with a diagnosis of sepsis," Medicina Intensiva, vol. 44, no. 3, pp. 160–170, 2020.
View at: Publisher Site | Google Scholar

5 J. Kim, H. Chang, D. Kim, D. H. Jang, I. Park, and K. Kim, "Machine learning for prediction of septic shock at initial triage in emergency department," Journal of Critical Care, vol. 55, pp. 163–170, 2020.
View at: Publisher Site | Google Scholar

6 A. H. Beck, A. R. Sangoi, S. Leung et al., "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," Science Translational Medicine, vol. 3, no. 108, 108ra113 pages, 2011.
View at: Publisher Site | Google Scholar

7 D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.
View at: Publisher Site | Google Scholar

8 R Core Team, R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, R Core Team, Vienna, Austria, 2014.

9 J. C. Gower, "A general coefficient of similarity and some of its properties," Biometrics, vol. 27, no. 4, pp. 857–871, 1971.
View at: Publisher Site | Google Scholar

10 T. Desautels, J. Calvert, J. Hoffman et al., "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach," JMIR Medical Informatics, vol. 4, no. 3, Article ID e5909, 2016.
View at: Publisher Site | Google Scholar

11 I. Taneja, B. Reddy, G. Damhorst, and S. D. Zhao, "Combining biomarkers with EMR data to identify patients in different phases of sepsis," Scientific Reports, vol. 7, no. 1, pp. 1–12, 2017.
View at: Publisher Site | Google Scholar

12 M. Scherpf, F. Gräßer, H. Malberg, and S. Zaunseder, "Predicting sepsis with a recurrent neural network using the MIMIC III database," Computers in Biology and Medicine, vol. 113, Article ID 103395, 2019.
View at: Publisher Site | Google Scholar

13 S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the ICU,"

Critical Care Medicine, vol. 46, no. 4, pp. 547–553, 2018.
View at: Publisher Site | Google Scholar

14 C. Lin, Y. Zhang, J. Ivy et al., "Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM," in Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 219–228, IEEE, New York, NY, USA, June 2018.

View at: Google Scholar

15 F. Lamping, T. Jack, N. Rübsamen et al., "Development and validation of a diagnostic model for early differentiation of sepsis and non-infectious SIRS in critically ill children-a data-driven approach using machine-learning algorithms," BMC Pediatrics, vol. 18, no. 1, pp. 1–11, 2018.
View at: Publisher Site | Google Scholar

16 S. Hariharan, Real-Time Sepsis Prediction Using an End-to-End Multi Task Gaussian Process RNN Classifier, Duke University, Durham, NC, USA, 2017.

17 M. A. Reyna, C. Josef, R. Jeter et al., "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," Critical care medicine, vol. 48, no. 2, pp. 210–217, 2019.
View at: Google Scholar

18 M. Reyna, C. Josef, R. Jeter et al., "Early prediction of sepsis from clinical data-the PhysioNet computing in cardiology challenge 2019 (version 1.0.0)," PhysioNet, 2019.
View at: Publisher Site | Google Scholar

19 J. O. Kim and J. Curry, "The treatment of missing data in multivariate analysis," Sociological Methods & Research, vol. 6, no. 2, pp. 215–240, 1977.
View at: Publisher Site | Google Scholar

20 Q. A. W. Raaijmakers, "Effectiveness of different missing data treatments in surveys with likert-type data: introducing the relative mean substitution approach," Educational and Psychological Measurement, vol. 59, no. 5, pp. 725–748, 1999.
View at: Publisher Site | Google Scholar

21 D. J. Stekhoven and P. Buhlmann, "MissForest—non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.
View at: Publisher Site | Google Scholar

22 P. Bühlmann and T. Hothorn, "Boosting algorithms: regularization, prediction and model fitting," Statistical Science, vol. 22, no. 4, pp. 477–505, 2007.
View at: Publisher Site | Google Scholar

23 Y. Song, X. Jiao, S. Yang et al., "Combining multiple factors of LightGBM and XGBoost algorithms to predict the morbidity of double-high disease," in Proceedings of the International Conference of Pioneering Computer Scientists, Engineers and Educators, pp. 635–644, Springer, Singapore, September 2019.
View at: Publisher Site | Google Scholar