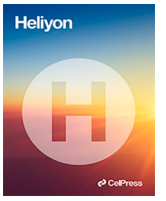
Heliyon 10 (2024) e29981

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Using machine learning algorithms based on patient admission laboratory parameters to predict adverse outcomes in COVID-19 patients

Yuchen Fu [a,b,1], Xuejing Xu [a,1], Juan Du [c], Taihong Huang [a], Jiping Shi [a], Guanghao Song [a], Qing Gu [b,***], Han Shen [a,**], Sen Wang [a,*]

[a] *Department of Clinical Laboratory Medicine, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, 210008, China*

[b] *State Key Laboratory for Novel Software Technology, National Institute of Healthcare Data Science at Nanjing University, Nanjing, 210008, China*

[c] *Comprehensive Cancer Center of Drum Tower Hospital, Medical School of Nanjing University, Clinical Cancer Institute of Nanjing University, Nanjing, 210008, China*

ARTICLE INFO

Keywords:
COVID-19
Machine learning
Lasso
SVM
Prognostic prediction

ABSTRACT

Amidst the global COVID-19 pandemic, the urgent need for timely and precise patient prognosis assessment underscores the significance of leveraging machine learning techniques. In this study, we present a novel predictive model centered on routine clinical laboratory test data to swiftly forecast patient survival outcomes upon admission. Our model integrates feature selection algorithms and binary classification algorithms, optimizing algorithmic selection through meticulous parameter control. Notably, we developed an algorithm coupling Lasso and SVM methodologies, achieving a remarkable area under the ROC curve of 0.9277 with the use of merely 8 clinical laboratory parameters collected upon admission. Our primary contribution lies in the utilization of straightforward laboratory parameters for prognostication, circumventing data processing intricacies, and furnishing clinicians with an expeditious and precise prognostic assessment tool.

1. Introduction

COVID-19, caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), emerged in Wuhan, China, in 2019, swiftly becoming a global pandemic with significant ramifications for society, economies, and healthcare systems [1]. The virus, primarily transmitted via respiratory droplets, exhibits a prolonged incubation period, facilitating transmission by asymptomatic carriers. Symptoms range from fever, cough, to severe manifestations like pneumonia, acute respiratory distress syndrome (ARDS), and multi-organ dysfunction [2]. Early identification of patients at risk of severe outcomes is imperative for effective intervention and treatment success.

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: guq@nju.edu.cn (Q. Gu), shenhan10366@sina.com (H. Shen), njwangsen@163.com (S. Wang).
[1] These authors contributed equally to this work and share first authorship.

https://doi.org/10.1016/j.heliyon.2024.e29981
Received 12 January 2024; Received in revised form 15 April 2024; Accepted 18 April 2024
Available online 21 April 2024

Numerous prognostic models for COVID-19 have been developed, relying on demographic data, comorbidities, physical exams, laboratory findings, and imaging [3–5]. Clinical laboratory parameters, including blood biochemistry markers, immunological parameters, and inflammatory biomarkers, offer crucial insights into prognosis. For instance, elevated inflammatory markers often signal disease exacerbation, while fluctuations in immunological parameters reflect the dynamic immune response.

To optimize the utility of clinical data and enhance predictive accuracy, integrating machine learning algorithms has become pivotal. Techniques like Support Vector Machines (SVM), Random Forest, and Logistic Regression can discern intricate relationships within data, facilitating precise predictions [6]. By amalgamating these algorithms with routine clinical data collected upon patient admission, predictive models furnish clinicians with timely and precise prognosis information, aiding in clinical decision-making and intervention strategies.

This study aims to develop a machine learning algorithm based on SVM, utilizing routine laboratory parameters collected upon patient admission for swift prognosis prediction in COVID-19 patients. The resultant algorithm serves as a reliable and expedient decision support tool for clinicians, enabling enhanced treatment and care for COVID-19 patients.

2. Methods

Our objective centers on prognosticating the outcomes of COVID-19 patients, specifically categorizing them into either the" fatal group" or" survival group", accomplished through a binary classification task. To delineate the overarching structure of our study, we depict the comprehensive framework in Fig. 1, characterized predominantly by three integral components: data preprocessing, feature selection, and binary classification.

2.1. Data source and study population

The data for this study originates from patients admitted to Nanjing Drum Tower Hospital of Jiangsu Province, China, who have all received a diagnosis of COVID-19 infections. The dataset consists of 599 patients, out of which 199 cases (33 %) exhibit varying degrees of missing clinical or experimental information. After excluding these 199 patients, a total of 400 patients were included in the final data analysis for this study.

Among the 400 patients, there are 275 male and 125 female patients. The age of the patients ranges from 30 to 104 years, with an average age of 76.69 years. The patient prognosis is distributed as follows: 330 cases have successfully recovered and regained their health, while 70 cases have unfortunately resulted in fatalities.

For the purpose of machine learning analysis, we have incorporated a comprehensive set of 63 indicators as features. These indicators encompass a range of clinical and experimental parameters, including but not limited to C-Reactive Protein (CRP), Estimated Glomerular Filtration Rate, and High-Density Lipoprotein Cholesterol (HDL-C).

2.2. Data preprocessing

In order to ensure the reliability and quality of our analysis, a series of meticulous data preprocessing steps were undertaken prior to conducting any subsequent analyses.

Initially, to maintain the integrity of laboratory parameters for feature selection, any samples with missing data were meticulously

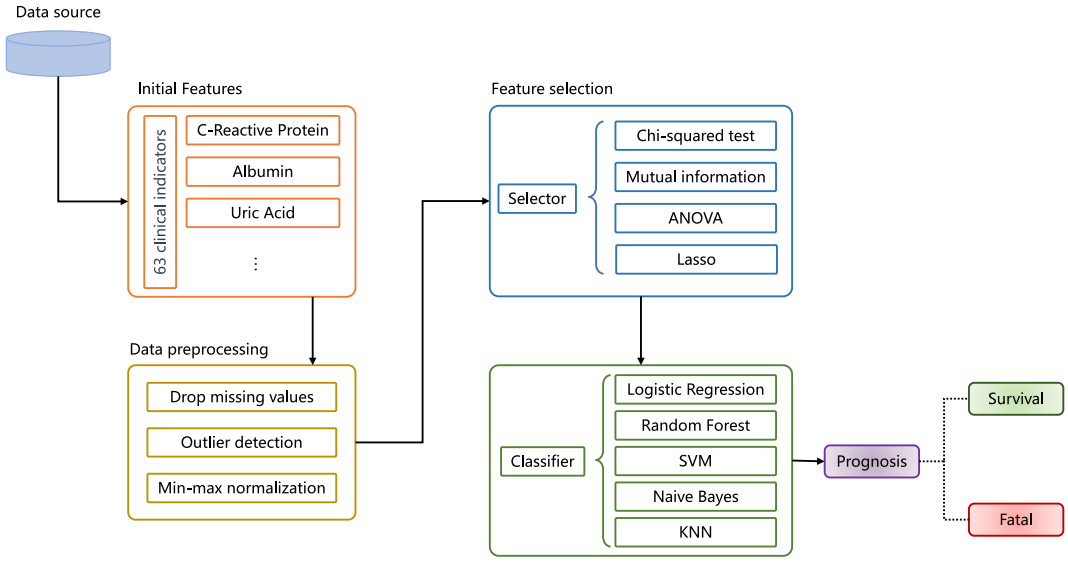


Fig. 1. The research framework of this study.

identified and subsequently excluded from the study. This careful step was undertaken to ensure that our subsequent analyses were based on a complete and accurate set of information. After addressing missing data, the dataset underwent a crucial outlier detection process. The z-score method was applied to pinpoint and eliminate outlier samples that exceeded a threshold of 3 times the standard deviation. By removing these extreme values, we aimed to mitigate the influence of potentially erroneous data points on our subsequent analyses. To enhance the comparability of the diverse feature values, a normalization procedure was meticulously executed. This step involved transforming the feature values of all samples to a standardized range, specifically within the interval of 0–1. By doing so, we aimed to mitigate any potential bias arising from the differing scales of individual features, thus enabling a more meaningful comparison and interpretation of results.

In essence, the data preprocessing phase constituted an essential cornerstone of our study. The exclusion of samples with missing data, identification and removal of outliers, and the subsequent normalization of feature values collectively laid a robust foundation for the subsequent machine learning analysis, ensuring the accuracy, reliability, and interpretability of our findings.

2.3. Machine learning algorithms

The core objective of our study is to predict patient prognosis mortality rates based on a comprehensive set of 63 features, employing binary classification machine learning algorithms. However, not all of the 63 features bear direct correlation to patient mortality rates. Following meticulous data preprocessing, we adopted established feature selection algorithms to refine and narrow down our feature set. Subsequently, we harnessed binary classification algorithms to forecast patient mortality rates, leveraging the chosen set of features.

2.3.1. Feature selection

To enhance interpretability, we opted for feature selection algorithms over dimensionality reduction methods like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Our study employs four distinct feature selection methods: the chi-squared test, mutual information, analysis of variance (ANOVA), and the Least Absolute Shrinkage and Selection Operator (Lasso).

The chi-squared test, designed to measure the association between two categorical variables, plays a pivotal role in feature selection. It assesses the relationship between a feature and the target variable by comparing observed and expected frequencies, quantifying the degree of association. This method is suitable for classification involving discrete variables [7].

Mutual information, a measure of dependence between random variables, evaluates the information sharing between a feature and the target variable. Higher values indicate stronger dependence, aiding in the selection of relevant features [8]. It works for both classification and regression tasks, handling both discrete and continuous features.

ANOVA, designed to analyze disparities among multiple groups, assesses variances between different groups, such as target variable categories [9]. By calculating the F-statistic, it determines if means across groups are equal, indicating a feature's potential influence. ANOVA is suited for classification scenarios with continuous features.

Lasso, with its L1 regularization component, excels in both feature selection and regression tasks. It forces certain feature coefficients to zero, effectively selecting important features. Lasso is valuable for datasets with many features and handles continuous features, making it efficient in scenarios with feature redundancy [10].

By integrating these feature selection methods, our analysis streamlines the feature space, enabling the identification of crucial predictors for patient prognosis mortality rates.

2.3.2. Binary classification algorithms

In our pursuit of predicting post COVID-19 infection patient survival, we carefully selected five binary classification algorithms: logistic regression, random forest, support vector machine (SVM), naive Bayes, and k-nearest neighbors.

(KNN). While our primary objective is to swiftly assess patients' prognoses based on laboratory indicators, we refrained from incorporating computationally intensive algorithms, such as neural networks.

Logistic regression, despite its name, is a vital linear classification algorithm. It maps the output of a linear function to a probabilistic realm before classifying based on a predefined threshold [11]. It excels in binary classification, producing a probabilistic estimate of class membership likelihood.

Random forest, an ensemble learning method, combines multiple decision trees, each trained on distinct data subsets. Their collective predictions, through voting or averaging, make classification decisions. Random forests shine in complex data scenarios with high-dimensional features [12].

SVM is a powerful classification algorithm aiming to find a hyperplane maximizing the separation between classes [13]. It handles non-linear data through kernel functions and identifies support vectors on class boundaries. SVM is effective in high-dimensional, small-sample datasets. Naive Bayes, based on Bayes' theorem, assumes feature independence. It works with discrete and continuous features, delivering robust performance even with limited data [14].

KNN classifies samples based on their proximity to training samples. It selects K nearest neighbors and determines the class through majority voting. KNN excels in uniform data distribution with limited noise [15]. Deploying these algorithms, our study reveals insights into patient survival prognosis post COVID-19 infection, leveraging their unique strengths to uncover complex patterns in our data.

3. Results

We conducted a comprehensive assessment of binary classification outcomes by employing established metrics including accuracy, Area Under the Curve (AUC), precision, recall, and F1 score. To ensure statistical robustness, all experimental results were acquired through the rigorous application of a 10-fold cross-validation methodology.

3.1. Feature selection performance

Fig. 2 displays the relationship between the percentage of selected features and the binary classification AUC scores. The range spans from 5 % to 100 %, with the SVM algorithm employed for binary classification. The graph indicates a gradual decline in binary classification performance as the number of features increases. Within the range of 10 %–15 % of the total feature count, optimal binary classification performance is observed, notably highlighted by the Lasso feature selection algorithm achieving an AUC score exceeding 0.925.

3.2. Feature importance

Table 1 details the subsets of features selected by the four feature selection algorithms at the point of highest SVM algorithm AUC. The final row of the table signifies the common features across the algorithms, elucidating shared critical features in this context.

We proceeded to perform an in-depth analysis of the contribution of the 8 features selected by the Lasso algorithm to the binary classification model, employing the SHapley Additive exPlanations (SHAP)method. This evaluation, as depicted in Fig. 3, provides insights into the significance and influence of each feature on the model's predictions.

3.3. Binary classification performance

A comprehensive comparative assessment of the five binary classification algorithms was conducted, as outlined in Table 2. All five algorithms utilized the feature subset derived from the Lasso algorithm as input, furnishing predictions regarding the survival outcomes of COVID-19 patients. Notably, our observations indicate that the recall of prediction outcomes generally surpasses precision, a phenomenon potentially attributed to the pronounced imbalance in sample composition between recovered and deceased cases.

Moreover, we observed that the Naive Bayes algorithm, while displaying a lower AUC compared to SVM, manifests superior scores in Accuracy, Precision, and F1 metrics. This divergence implies the susceptibility of the Naive Bayes approach to the influence of class imbalance, rendering it more adept at handling such scenarios.

Fig. 4a exhibits the Receiver Operating Characteristic (ROC) curves corresponding to each fold within the context of the 10-fold cross-validation. Meanwhile, Fig. 4b visually presents the composite view of the average ROC curves for the suite of five distinct binary classification algorithms.

The intricacies of classification model reliability are demonstrated through the reliability curves [16] depicted in Fig. 5. This graphical representation conveys the alignment between predicted probabilities generated by the classification model and the actual observed frequencies. The extent of deviation from the 45-degree diagonal line indicates the discrepancy between model predictions and observed outcomes. Notably, the reliability curves underscore the heightened predictive accuracy and confidence exhibited by logistic regression and SVM models.

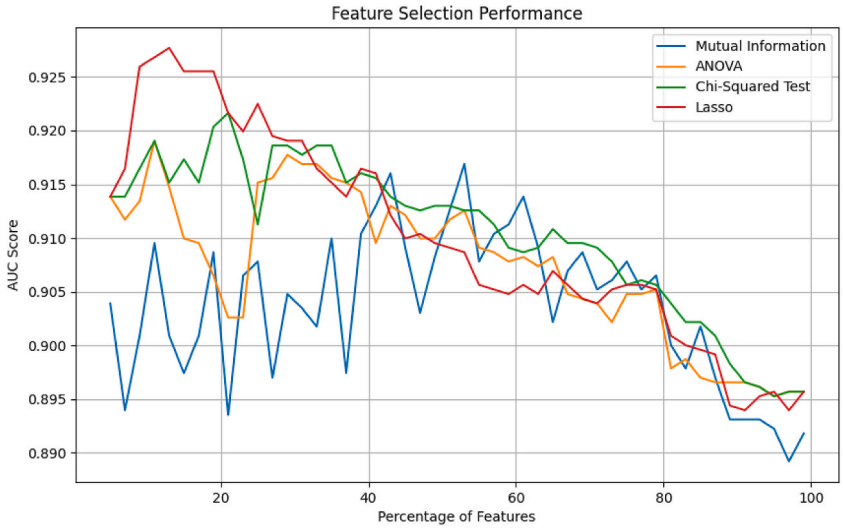


Fig. 2. Percentage of features used for training versus AUC score for determining the optimal number of features using the SVM.

Table 1
Best performing feature sets for feature selection algorithms.

Method	number of features	AUC	Features
Chi-Squared	13	0.9216	C-Reactive Protein, Cholinesterase, Triglycerides, Aspartate Aminotransferase (AST), Creatinine, Blood urea nitrogen (BUN), Lactate Dehydrogenase (LDH), Direct Bilirubin, Lymphocyte Percentage, Basophil Percentage, Atypical Lymphocyte Percentage, Nucleated Red Blood Cell Count (NRBC Count), D-Dimer
Mutual Information	9	0.9229	C-Reactive Protein, High-Density Lipoprotein Cholesterol (HDL-C), Gamma-Glutamyl Transpeptidase (GGT), Alkaline Phosphatase (ALP), Blood urea nitrogen (BUN), Lactate Dehydrogenase (LDH), Apolipoprotein A1 (ApoA1), Direct Bilirubin, Hematocrit (Hct), Lymphocyte Percentage, Lymphocyte Count, Nucleated Red Blood Cell Count (NRBC Count), D-Dimer
ANOVA	6	0.9190	C-Reactive Protein, Cholinesterase, Aspartate Aminotransferase (AST), Blood urea nitrogen (BUN), Lactate Dehydrogenase (LDH), Direct Bilirubin
Lasso	8	0.9277	Lactate Dehydrogenase (LDH), C-Reactive Protein, Blood urea nitrogen (BUN), Atypical Lymphocyte Percentage, Cholinesterase, Platelet Count, Monocyte Percentage, Total Bilirubin
Intersection	3	/	Blood urea nitrogen (BUN), Lactate Dehydrogenase (LDH), C-Reactive Protein

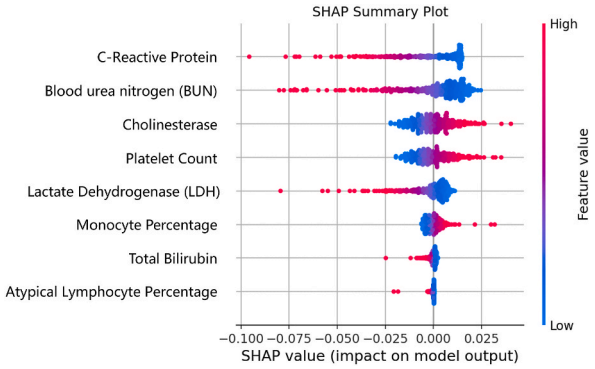


Fig. 3. SHAP-based feature importance.

Table 2
Performance results of the five binary classification algorithms.

Classifier	Accuracy	AUC	Precision	Recall	F1
LogisticRegression	0.8250	0.9160	0.8250	**0.9999**	0.9041
RandomForest	0.8675	0.9030	0.8747	0.9818	0.9246
SVM	0.8325	**0.9277**	0.8347	0.9939	0.9073
Naive Bayes	**0.8800**	0.9009	**0.9301**	0.9273	**0.9271**
KNN	0.8750	0.8816	0.9022	0.9545	0.9268

Lastly, we present the results of dimensionality reduction using Principal Component Analysis (PCA) [17] and t-distributed Stochastic Neighbor Embedding (t-SNE) [18] algorithms in Fig. 6a and b. Employing these techniques, the 8 features selected by the Lasso algorithm were transformed into two dimensions, and their distribution was visualized. This visualization reveals distinct distribution patterns between recovered and deceased cases, represented by yellow and purple dots respectively, highlighting the discernible dissimilarity between the two groups.

4. Discussion

The global outbreak of the COVID-19 pandemic has made timely and effective assessment of patient prognosis an urgent necessity. Machine learning holds significant importance in predicting the prognosis of COVID-19 patients. Machine learning leverages large-scale clinical data to uncover potential factors related to prognosis. It establishes predictive models, assisting doctors in formulating better individualized treatment and care plans. Consequently, this enhances patients' survival rates and the quality of their recovery. Currently, there are several machine learning models employed for predicting the prognosis of COVID-19 patients. However, most of these models rely on a diverse range of complex data such as clinical symptoms and medical imaging, resulting in the intricacy of data collection and processing, as well as time-consuming prediction processes [5,19]. Here, we present a prognosis prediction model based on routine clinical laboratory test data. By utilizing routine clinical laboratory parameters collected upon patients' admission, rapid prognosis prediction for patients is achieved without the need for complex data. We predict patient mortality by combining feature selection algorithms with binary classification algorithms, and we select the optimal algorithm combination

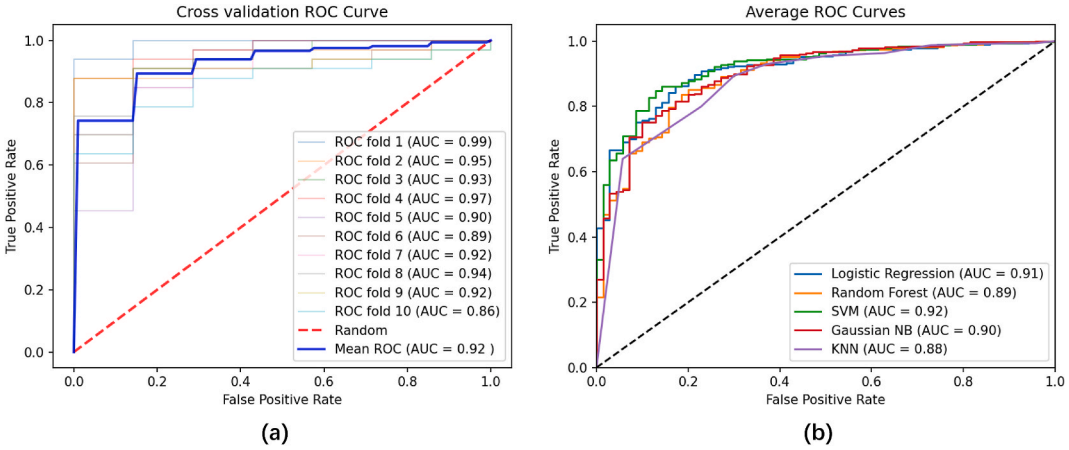


Fig. 4. ROC curves for binary classification models predicting COVID-19 survival outcomes. (a) ROC curves for each fold in 10-fold cross-validation. (b) ROC curves for 5 binary classification algorithms.

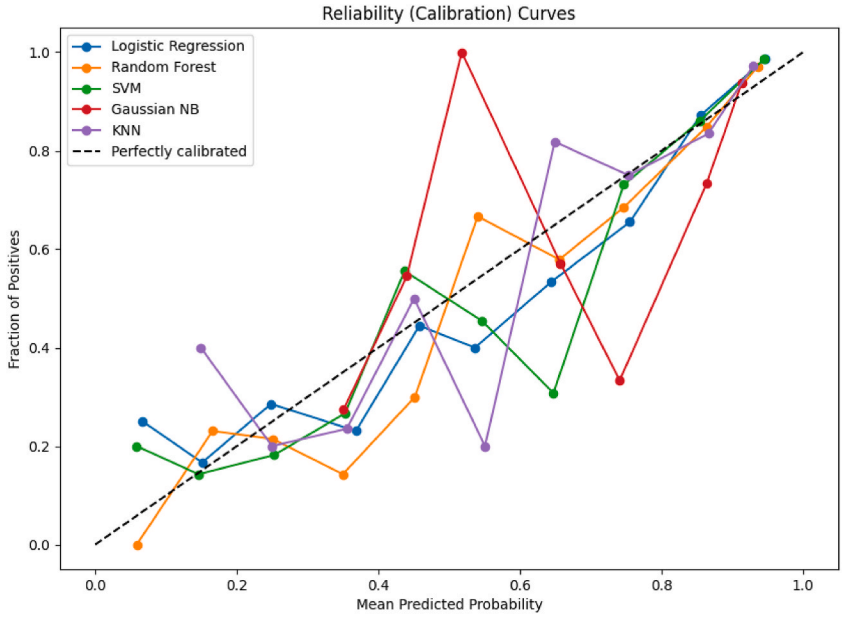


Fig. 5. Reliability curves for the five binary classification algorithms.

through controlled variable methods. Ultimately, we have constructed an algorithm based on Lasso and SVM, utilizing eight clinical laboratory test data points obtained upon patient admission: LDH, CRP, BUN, Atypical Lymphocyte Percentage, Cholinesterase, Platelet Count, Monocyte Percentage, and Total Bilirubin. This algorithm accurately predicts patient prognosis, with an area under the ROC curve reaching 0.9277.

The culminating experimental outcomes substantiate the superior performance of the amalgamation involving Lasso feature selection and the SVM binary classification algorithm. This success can be attributed to three pivotal factors. First, the Lasso algorithm adeptly distills crucial features even in the presence of multicollinearity [20]. Second, the SVM algorithm exhibits a robust resistance to challenges stemming from class imbalance [21], its intrinsic ability to recalibrate penalty terms contributes to a pronounced focus on the minority class, thereby heightening predictive accuracy. Third, the convergence of Lasso and SVM, both entrenched in the realm of linear models, underscores their harmonious synergy and consequent compatibility [22].

CRP, LDH, and BUN were identified as common features selected by several models in our investigation. These clinical parameters have also been demonstrated to be associated with patient prognosis in multiple studies. CRP is a non-specific inflammatory marker, with a significant increase observed in CRP levels among the majority of severe patients [23]. Furthermore, meta-analyses have indicated a notable correlation between elevated CRP levels and the severity of COVID-19 in patients [24]. LDH is commonly present in various tissues, and SARS-CoV-2 can directly infect lung cells, leading to tissue damage and subsequently causing an increase in LDH

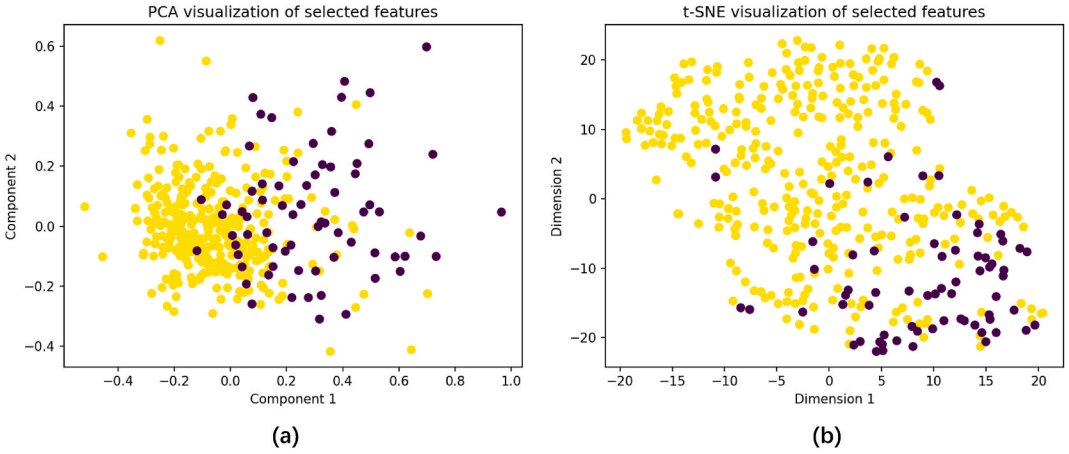


Fig. 6. Visualization of features selected by Lasso using PCA and t-SNE algorithms. (a) PCA. (b) t-SNE.

levels. The inflammatory response triggered by SARS-CoV-2 can also contribute to elevated LDH levels. Elevated LDH levels can raise the likelihood of severe illness in patients by approximately 6 times and increase the mortality rate by about 16 times [25]. Serum BUN, an essential indicator of kidney function, plays a critical role in understanding COVID-19 progression and mortality, which are closely associated with multi-organ dysfunction. Kidney impairment in COVID-19 patients may result from SARS-CoV-2's direct attack, cytokine storms, and hypoxemia. Furthermore, a systematic review has highlighted elevated serum BUN levels as a key risk factor for increased severity and mortality in these patients [26].

In this study, the Lasso algorithm demonstrated the highest AUC, with its selected features comprising not only CRP, LDH, and BUN but also five additional laboratory parameters. Thrombocytopenia is a significant feature of SARS-CoV-2 infection and an important indicator of poor prognosis. Increased platelet consumption and decreased production may be contributing factors to the development of thrombocytopenia [27,28]. Cholinesterase is an enzyme produced in the liver that hydrolyzes acetylcholine. Impairment of liver function and the influence of inflammatory cytokines may result in a reduction of cholinesterase levels. The cholinesterase level upon patient admission can serve as one of the predictive factors for severity and prognosis [29]. The elevation of total bilirubin reflects damage to the patient's liver function, and severe impairment of liver function is also a significant influencing factor for adverse patient prognosis. The counts of atypical lymphocytes and monocytes are two hematological parameters identified by our model. The increase in atypical lymphocytes and monocytes could be attributed to SARS-CoV-2 infection, and these abnormalities are crucial characteristics during the early stages of COVID-19 infection [30]. The feature parameters identified by the machine learning model constructed in this study exhibit significant variations during the course of COVID-19 infection. Our model achieves precise prediction of patient prognosis by integrating these crucial laboratory parameters, enhancing the accuracy of prognosis assessment.

Our dataset is limited by its relatively small sample size, which consists solely of patients from a single hospital. As a result, the scalability of our proposed method to larger datasets may potentially incur performance degradation. Furthermore, our machine learning approach is designed as a two-stage process, involving the selection of features and classification algorithms, both of which can significantly impact outcomes. This could lead to suboptimal results, especially in scenarios where significant disparities exist between the distributions of test and training data. Moving forward, we aim to explore end-to-end deep learning methodologies and validate our approach using larger-scale, multicenter datasets. This will enhance the generalizability and practical utility of our method.

In summary, we have designed an algorithm that integrates Lasso and SVM techniques, achieving precise prognostic predictions for COVID-19 patients based on just 8 clinical laboratory parameters collected early upon hospital admission. The strength of our study lies in its use of straightforward clinical laboratory parameters for prediction, thereby simplifying data processing and offering clinicians a rapid and accurate tool for prognostic assessment.

Ethical approval statement

This study was approved by the Institutional Review Board (IRB) of Nanjing Drum Tower Hospital (2022–746), Nanjing, China. Due to the retrospective nature of the study, the ethics committee (IRB of Nanjing Drum Tower Hospital) waived the requirement for patient consent.

Consent for publication

Not applicable.

Funding

This work was supported by grants from the National Key Research and Development Program of China (2023YFC2309100), Clinical Trials from the Affiliated Drum Tower Hospital, Medical School of Nanjing University (2022-LCYJ-MS-28, 2022-LCYJ-PY-40) and Nanjing Medical Science and technique Development Foundation (ZKX23023, QRX17142, YKK21066), The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

CRediT authorship contribution statement

Yuchen Fu: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Xuejing Xu:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Juan Du:** Formal analysis, Data curation. **Taihong Huang:** Formal analysis, Data curation. **Jiping Shi:** Data curation. **Guanghao Song:** Data curation. **Qing Gu:** Writing – review & editing, Validation, Supervision, Software, Resources, Conceptualization. **Han Shen:** Writing – review & editing, Validation, Supervision, Software, Resources, Conceptualization. **Sen Wang:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not applicable.

References

[1] Chen Wang, Peter W. Horby, Frederick G. Hayden, George F. Gao, A novel coronavirus outbreak of global health concern, The lancet 395 (10223) (2020) 470–473.
[2] Kunyu Yang, Yuhan Sheng, Chaolin Huang, Yang Jin, Nian Xiong, Ke Jiang, Hongda Lu, Jing Liu, Jiyuan Yang, Youhong Dong, et al., Clinical characteristics, outcomes, and risk factors for mortality in patients with cancer and covid-19 in hubei, China: a multicentre, retrospective, cohort study, Lancet Oncol. 21 (7) (2020) 904–913.
[3] Laure Wynants, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Elena Albu, Banafsheh Arshi, Vanesa Bellou, Marc MJ. Bonten, et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ 369 (2020).
[4] Mahdieh Montazeri, Roxana ZahediNasab, Farahani Ali, Hadis Mohseni, Fahimeh Ghasemian, Machine learning models for image-based diagnosis and prognosis of covid-19: systematic review, JMIR medical informatics 9 (4) (2021) e25181.
[5] Rufaidah Dabbagh, Amr Jamal, Jakir Hossain Bhuiyan Masud, Maher A. Titi, Yasser S. Amer, Afnan Khayat, Taha S. Alhazmi, Layal Hneiny, Fatmah A. Baothman, Alkubeyyer Metab, et al., Harnessing machine learning in early covid-19 detection and prognosis: a comprehensive systematic review, Cureus 15 (5) (2023).
[6] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.
[7] Anna Karen Garate-Escamila, Amir Hajjam El Hassani, Emmanuel Andres, Classification models for heart disease prediction using feature selection and pca, Inform. Med. Unlocked 19 (2020) 100330.
[8] Hongfang Zhou, Xiqian Wang, Rourou Zhu, Feature selection based on mutual information with correlation coefficient, Appl. Intell. 1–18 (2022).
[9] Bertinetto Carlo, Jasper Engel, Jeroen Jansen, Anova simultaneous component analysis: a tutorial review, Anal. Chim. Acta 6 (X) (2020) 100061.
[10] Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim, FM Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Ab- hijith Reddy Beeravolu, Friso De Boer, Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques, IEEE Access 9 (2021) 19304–19326.
[11] Simon Nusinovici, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, Ching- Yu Cheng, Logistic regression was as good as machine learning for predieting major chronic diseases, Journal of clinical epidemiology 122 (2020) 56–69.
[12] Prajyot Palimkar, Rabindra Nath Shaw, Ankush Ghosh, Machine learning technique to prognosis diabetes disease: random forest classifier approach, in: Advanced Computing and Intelligent Technologies: Proceedings ofICACIT 2021, Springer, 2022, pp. 219–244.
[13] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodriguez-Mazahua, As- drubal Lopez, A comprehensive survey on support vector machine classification: applications, challenges and trends, Neurocomputing 408 (2020) 189–215.
[14] Indika Wickramasinghe, Harsha Kalutarage, Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation, Soft Comput. 25 (3) (2021) 2277–2293.
[15] Shler Farhad Khorshid and Adnan Mohsin Abdulazeez, Breast cancer diagnosis based on k-nearest neighbors: a review, PalArch's Journal of Archaeology of Egypt/Egyptology 18 (4) (2021) 1927–1951.
[16] Seyed Mojtaba Moosavi, Sussan Ghassabian, Linearity of calibration curves for analytical methods: a review of criteria for assessment of method reliability, in: Calibration and Validation of Analytical Methods: a Sampling of Current Approaches, vol. 109, 2018.
[17] Caterina Labrin, Francisco Urdinez, Principal component analysis, in: R for Political Data Science, Chapman and Hall/CRC, 2020, pp. 375–393.
[18] Matthew C. Cieslak, Ann M. Castelfranco, Vittoria Roncalli, Petra H. Lenz, Daniel K. Hartline, t-distributed stochastic neighbor embedding (t-sne): a tool for eco-physiological transcriptomic analysis, Mar. Genomics 51 (2020) 100723.
[19] Ruiyao Chen, Jiayuan Chen, Sen Yang, Shuqing Luo, Zhongzhou Xiao, Lu Lu, Bilin Liang, Sichen Liu, Huwei Shi, Jie Xu, Prediction of prognosis in covid-19 patients using machine learning: a systematic review and meta-analysis, Int. J. Med. Inf. 105151 (2023).

[20] Shady Altelbany, Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: a simulation study, Journal of Applied Economics and Business Studies 5 (1) (2021) 131–142.
[21] Rehan Akbani, Stephen Kwek, Nathalie Japkowicz, Applying support vector machines to imbalanced datasets, in: Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15, Springer, 2004, pp. 39–50.
[22] Chuanze Kang, Yanhao Huo, Lihui Xin, Baoguang Tian, Bin Yu, Feature selection and tumor classification for microarray data using relaxed lasso and generalized multi-class support vector machine, J. Theor. Biol. 463 (2019) 77–91.
[23] Giuseppe Lippi, Mario Plebani, Laboratory abnormalities in patients with covid-2019 infection, Clin. Chem. Lab. Med. 58 (7) (2020) 1131–1134.
[24] Preeti Malik, Urvish Patel, Deep Mehta, Nidhi Patel, Raveena Kelkar, Muhammad Akrmah, Janice L. Gabrilove, Sacks Henry, Biomarkers and outcomes of covid-19 hospitalisations: systematic review and metaanalysis, BMJ evidence-based medicine 26 (3) (2021) 107–108.
[25] Brandon Michael Henry, Gaurav Aggarwal, Johnny Wong, Stefanie Benoit, Jens Vikse, Mario Plebani, Giuseppe Lippi, Lactate dehydrogenase levels predict coronavirus disease 2019 (covid-19) severity and mortality: a pooled analysis, The American journal of emergency medicine 38 (9) (2020) 1722–1726.
[26] Ariel Izcovich, Martin Alberto Ragusa, Fernando Tortosa, Maria Andrea Lavena Marzio, Camila Agnoletti, Agustin Bengolea, Agustina Ceirano, Federico Espinosa, Ezequiel Saavedra, Veronica Sanguine, et al., Prognostic factors for severity and mortality in patients infected with covid-19: a systematic review, PLoS One 15 (11) (2020) e0241955.
[27] Yanli Liu, Wenwu Sun, Yanan Guo, Liangkai Chen, Lijuan Zhang, Su Zhao, Ding Long, Li Yu, Association between platelet parameters and mortality in coronavirus disease 2019: retrospective cohort study, Platelets 31 (4) (2020) 490–496.
[28] Giuseppe Lippi, Mario Plebani, Brandon Michael Henry, Thrombocytopenia is associated with severe coronavirus disease 2019 (covid-19) infections: a meta-analysis, Clinica chimica acta 506 (2020) 145–148.
[29] Kento Nakajima, Takeru Abe, Ryo Saji, Fumihiro Ogawa, Hayato Taniguchi, Keishi Yamaguchi, Kazuya Sakai, Tomoki Nakagawa, Reo Matsumura, Yasufumi Oi, et al., Serum cholinesterase associated with covid-19 pneumonia severity and mortality, J. Infect. 82 (2) (2021) 282–327.
[30] Andrea Lombardi, Elena Trombetta, Alessandra Cattaneo, Valeria Castelli, Emanuele Palomba, Mario Tirone, Davide Mangioni, Giuseppe Lamorte, Maria Manunta, Daniele Prati, et al., Early phases of covid-19 are characterized by a reduction in lymphocyte populations and the presence of atypical monocytes, Front. Immunol. 11 (2020) 560330.