Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2026, IX (4): 6756 – 6763 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

PREDICTING MORTALITY FOR COVID-19 PATIENTS USING DATA MINING AND MACHINE LEARNING APPROACHES

Elvin Khoirunnisa[1]*, Jin Xu[2]
[1*]Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
[2]INESC TEC and University of Porto, Portugal
email: *[1]elvinkhoirunnisa@gmail.com, [2]jinxu130817@gmail.com *corresponding author

Abstract: *Since December 2019, the COVID-19 pandemic has spread worldwide, posing a grave danger to global public health and severely threatening healthcare systems. Therefore, early prediction of COVID-19 case severity is crucial for saving lives and optimizing healthcare logistics. In this study, we utilize a blood sample dataset from infected patients in Wuhan, China, and construct machine learning models to predict patient mortality risk. Given that the testing data includes only three biomarkers, the experiments are divided into two parts: an "all features" scenario and a "three features" scenario. Based on feature selection and model optimization, candidate models are configured with different feature subsets and evaluated using both single algorithms and ensemble models on the training data. The trained models are then evaluated on the testing data. The results demonstrate that ensemble models achieve superior performance compared to single models across both feature scenarios. Notably, ensemble models using the "three features" configuration achieve the highest overall performance among all experimental setups, despite relying on fewer features. Our top F1 score for the positive class (death) approaches the baseline results, confirming the effectiveness of our experimental framework.*

Keywords: *Predictive model, COVID-19 mortality, data mining, feature selection*

Abstrak: Sejak Desember 2019, wabah COVID-19 telah menyebar ke seluruh dunia, tidak hanya membahayakan nyawa dan kesehatan penduduk global serta mengancam sistem layanan kesehatan masyarakat secara serius. Oleh karena itu, prediksi dini mengenai tingkat keparahan kasus COVID-19 sangatlah penting untuk kehidupan masyarakat dan perencanaan logistik kesehatan. Dalam penelitian ini, kami memanfaatkan basis data sampel darah dari pasien yang terinfeksi di wilayah Wuhan, Tiongkok, dan membangun model pembelajaran mesin (*machine learning*) untuk memprediksi risiko kematian pasien tersebut. Mengingat data pengujian hanya mencakup tiga biomarker, eksperimen ini dibagi menjadi dua bagian: skenario "semua fitur" dan "tiga fitur". Berdasarkan pemilihan fitur dan model, berbagai model dikonfigurasi dengan rangkaian fitur yang berbeda, lalu menerapkan model tunggal maupun model *ensemble* pada data pelatihan. Selanjutnya, model yang telah dilatih tersebut diuji menggunakan data pengujian. Hasil penelitian menunjukkan bahwa model *ensemble* mencapai kinerja yang lebih baik dibandingkan model tunggal, baik pada skenario "semua fitur" maupun "tiga fitur". Hasil dari model *ensemble* pada skenario "tiga fitur" merupakan yang tertinggi di antara seluruh rangkaian eksperimen, meskipun menggunakan jumlah fitur yang lebih sedikit. Skor F1 tertinggi kami untuk sampel positif (*death*) mendekati hasil eksperimen dasar (*baseline*), yang menunjukkan efektivitas pendekatan eksperimental dalam penelitian ini.

Kata kunci: Model prediktif, mortalitas COVID-19, penambangan data, pemilihan fitur

INTRODUCTION

Since the initial outbreak of novel coronavirus in Wuhan, China, in

December 2019, the exponential transmission of COVID-19 has severely disrupted daily life and placed immense pressure on global healthcare infrastructure, driving cumulative cases up by over three million in early 2020 alone (Sarwar et al., 2021) . This rapid escalation created critical shortages in intensive care resources, underscoring the urgent need for early risk stratification to facilitate strategic logistical planning, such as optimizing medical staffing and physical inventory, while prioritizing high-risk individuals to help reduce mortality. To address this challenge, this study leverages data science and machine learning techniques to construct accurate mortality prediction models using blood panel data from infected patients at Tongji Hospital in Wuhan, China, building upon the baseline framework established by Yan et al., (2020). Machine learning means the capability of systems to learn from specific training data related to a problem, enabling them to automate the development of analytical models and effectively solve the tasks involved (Janiesch et al., 2021).

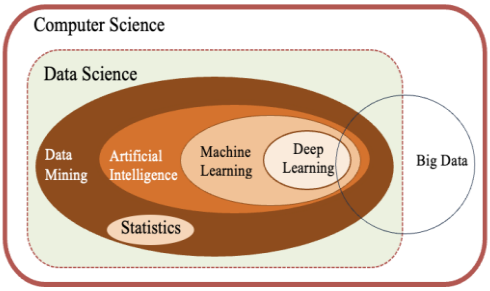

Figure 1. A Venn diagram illustrates the connections between methods based on computer science. (Source: Gordan et al., (2022))

Prior research has also explored COVID-19 patient mortality prediction using the broad learning system (R. Han et al., 2020), data mining techniques (Moulaei, 2021; Shanbehzadeh et al., 2021), the random forest algorithm with different datasets in Hungarian ICU settings (Hamar et al., 2024), artificial intelligence frameworks (Halwani & Halwani, 2024), time series and machine learning methods (Akter et al., 2024; Yildirim et al., 2024), and supervised learning models for (Khuluq et al., 2024).

What sets our study apart is our systematic evaluation of these pipeline decisions. Specifically, we evaluate how missing data imputation strategies, feature normalization, and feature subset selection (comparing a full feature set against a minimal three-biomarker set) directly influence the predictive performance of both single algorithms and ensemble models.

Specifically, this study addresses two core research questions:

a. What are the features that influence the prediction of COVID-19 mortality?
b. What model does have the best prediction for COVID-19 mortality*?*

To answer these questions, our experimental pipeline incorporates data normalization (scaling feature values to a 0–1 range), cross-validation for feature and model selection, and performance evaluation across two distinct input configurations, a restricted three-biomarker panel versus an expanded set containing all available features that will be explained in next sections.

METHOD

This research adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, utilizing five of its six core phases (see Figure 2): Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. As illustrated in Figure 3, these phases are integrated with the study's research questions to establish a structured methodology for predicting COVID-19 patient outcomes ("death" or "discharge").

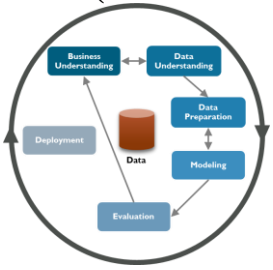

Figure 2. CRISP-DM Model (Source: dmlab.hu)

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2026, IX (4): 6756 – 6763 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

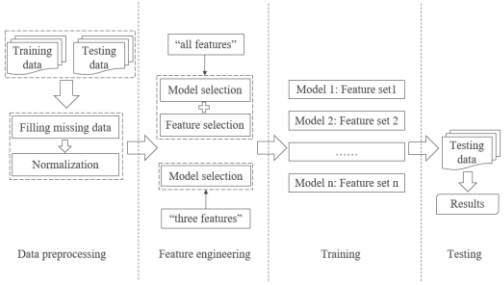

Figure 3. The Framework of Building Mortality Prediction Model for COVID-19 Patients

The experimental workflow begins with data preprocessing applied to both the training and testing sets, comprising missing value imputation followed by feature normalization to standard scales. The experiments are then divided into two parts: an "all features" setup and a "three features" setup. In the full feature setup, we run feature selection alongside model selection to identify the strongest feature-model pairs. In the three-feature setup, we focus solely on model selection. Once configured, each model is trained on its assigned feature set and assessed on the testing data.

1) Business Understanding

The goal of this phase is to define the project's core objective: building a prediction model for COVID-19 mortality based on clinical blood data. When provided with a patient's blood sample measurements, the model classifies the outcome into a binary format, where **0** indicates patient discharge and **1** indicates mortality.

2) Data Understanding

The dataset consists of training and testing data from infected patients at Tongji Hospital in Wuhan (Yan et al., 2020).

For training data, it has 74 features, and testing data has only 3 features (LDH, lymphocytes, and hs-CRP). This study uses a dataset split into training and testing sets. The training data contains 74 clinical features, whereas the testing data includes only three features: lactate dehydrogenase (LDH), lymphocytes, and high-sensitivity C-reactive protein (hs-CRP).

The dataset provides multiple medical records per patient, reflecting their health status at different points in time. All available records are retained in this study, as each measurement captures valuable information about the patient's changing condition, regardless of whether the final outcome remain the same.

3) Data Preprocessing

Based on an initial inspection of the training and testing datasets, several data quality issues were identified and addressed using the following preprocessing methods.

First, both datasets contained a significant amount of missing data, requiring imputation across the majority of records. Because the optimal imputation strategy was not known *a priori*, we tested four standard substitution techniques (Table 1). Each method was evaluated downstream, and the one yielding the highest model performance was selected.

Table 1. The Inserted Values for Missing Data

Inserted values	Description
0	All the missing records are filled in with 0.
Mean	For each column, the missing records are filled in with the average of that column.
Median	For each column, the missing records are filled in with the median of that column.
Mode	For each column, the missing records are filled in with the mode of that column.

Second, feature values varied significantly in scale. In machine learning, unscaled features with large numerical ranges can dominate model optimization, leading to biased predictions. For instance, in our dataset, LDH values ranged from 119 to 1867, whereas hs-CRP values ranged from 0.1 to 317.6. To prevent high-magnitude features from disproportionately influencing model

parameters, min-max normalization was applied using Equation (1):

$$x_{new} = \frac{x_{old} - min}{max - min} \quad (1)$$

Min and max mean the minimum and maximum value of the column in which x is located. After data transformation with normalization, values in each column range from 0 to 1.

Finally, because the official testing set contains only three biomarkers (LDH, lymphocytes, and hs-CRP), we designed a two-pronged experimental structure to evaluate performance under both restricted and full feature conditions:

a. **All Features Configuration:** To test models trained on a broader set of biomarkers, we shuffled the original training dataset and performed a stratified split, reserving 33% for testing and using the remaining 67% for training. This stratification ensured that both subsets maintained identical proportions of positive (death) and negative (discharge) outcomes.
b. **Three Features Configuration:** For direct evaluation against the external test set, we restricted the dataset exclusively to the three available biomarkers (LDH, lymphocytes, and hs-CRP), maintaining identical features across both training and testing phases.

4) Modeling

Regarding the tools we used in conducting our experiments, we mainly used Python as our programming language(Unpingco, 2016). In the process of doing data preprocessing, NumPy[1] was utilized to read the data from Excel into memory, storing them in DataFrame format with pandas[2]. When training models, libraries available in sklearn[3] like SVM and Decision Tree were applied to fitting the dataset. About the evaluation of the model, we also used sklearn to obtain the confusion matrix since our research question is regarded as a two-class classification problem.

Considering that the ensemble model tended to achieve better performance because of its complex and robust structures, we used single and ensemble models to train the data (see Table 2). For these two types of models, feature selection and model were applied to obtain optimal model-feature combinations. Therefore, we firstly conducted feature selection and model selection on training data. Then, features showed different performances in different models. Therefore, based on the results from cross-validation, we configured different models with different feature sets.

Table 2. Models of Single and Ensemble Categories

Category	Models
Single models	KNN3, KNN5, KNN7, SVM(kernel=rbf,poly,sigmoid), NB, DT, LR
Ensemble models	RF5, RF10, RF15, AdaBoost(estimators=100), XGBoost, GradientBoost

KNN3:K-Nearest Neighbor(neighbours=3); Support Vector Machines (SVM); NB: Naive Bayes; DT: Decision Tree; LR: Logistic Regression; RF5: Random Forest (trees=5)

5) Evaluation

In clinical decision-making, classification errors carry asymmetric risks. Misclassifying a high-risk patient as low-risk (a false negative for mortality) can delay critical care and lead to preventable death. Conversely, misclassifying a recovering patient as high-risk (a false positive) leads to unnecessary diagnostic tests, prolonged hospitalization, and resource strain, but does not carry immediate mortality risk.

Consequently, accurately identifying patients at risk of mortality (the positive class) is far more critical than

[1] https://numpy.org/
[2] https://pandas.pydata.org/

identifying patients likely to be discharged (the negative class). For this reason, the **F1 score for the positive class (death)** was selected as the primary evaluation metric, as it balances both precision and recall for high-risk predictions. Additional metrics, including overall precision, recall, and negative-class performance, are made available in the project's public GitHub repository.

RESULTS AND DISCUSSION

Feature selection and model selection

a. "all features" part

Based on the cross-validation of training data, we selected the features with much higher F1 scores on positive samples. We conducted this with different data-filling methods on unnormalized and normalized data.

Table 3. **Feature and Model Selection for Single Models with Unnormalized and Normalized Data in "all features" part**

Unnormalized (median)		**Normalized (mode)**	
KNN3	67 features	KNN3	54 features
KNN5	68 features	KNN5	52 features
KNN7	69 features	KNN7	53 features
NB	7 features	NB	14 features
SVM(sigmoid)	1 feature	SVM(sigmoid)	3 features

Based on the positive-class F1 scores, we selected the optimal feature combinations for each candidate model. As shown in Table 3, median imputation yielded the best performance for unnormalized data, whereas mode imputation proved most effective for normalized data. Consequently, individual model architectures were configured with their respective top-performing feature subsets.

Table 4. **Feature and Model Selection for Ensemble models with Unnormalized and Normalized Data in "all features" part**

Unnormalized (0)		**Normalized (0)**	
RF5	39 features	RF5	38 features
RF10	36 features	RF10	37 features
RF15	39 features	RF15	40 features
AdaBoost	37 features	AdaBoost	38 features
Gradient Boost	38 features	Gradient Boost	38 features
XGBoost	36 features	XGBoost	39 features

A similar selection process was applied to the ensemble architectures to determine their optimal feature configurations (Table 4). Interestingly, zero-imputation 0 consistently produced superior results for ensemble models across both normalized and unnormalized datasets, highlighting a distinct preference in data preparation between single and ensemble learning approaches.

b. "three features" part

Since there are only three features, we omitted feature selection and only conducted model selection for the "three features" part. The results from cross-validations on the training dataset are given in Table 5 and Table 6 below.

Table 5. **Model Selection for Single Models with Unnormalized and Normalized Data in "three features"**

Unnormalized (mode)	**Normalized (mode)**
KNN3, KNN5, KNN7, NB, DT, LG	KNN3, KNN5, KNN7, NB, DT, LG, SVM(rbf), SVM(poly), SVM(sigmoid)

Table 6. **Model Selection for Ensemble Models with Unnormalized and Normalized Data in "three features"**

Unnormalized (0)	Normalized (mode)
RF5, RF10, RF15, AdaBoost, GradientBoost, XGBoost	RF5, RF10, RF15, AdaBoost, GradientBoost, XGBoost

1) Results on the testing dataset

This section presents the results by testing the single models and ensemble models, which were trained on the training data.

a. "all features" part

For the "all features" part, every single model and ensemble model is configured with their best-performance features sets and applied to predicting the mortality of COVID-19 patients in testing data.

Table 7. **Performance of Single models in "all features"**

Model	Unnormalized	Normalized
KNN3	0,7442	0,7343
KNN5	0,7475	0,7620
KNN7	0,7489	0,7669
NB	0,6208	0,6393
SVM(sigmoid)	0,6079	0,5832

As indicated in Table 7, data normalization did not yield a significant performance improvement for single-algorithm models. In fact, for 3-Nearest Neighbors (KNN3) and SVM-sigmoid, performance on normalized data was slightly lower than on unnormalized data.

Table 8. **Performance of Ensemble Models in "all features" part**

Model	Unnormalized	Normalized
RF5	0,7262	0,7063
RF10	0,7044	0,7057
RF15	0,7404	0,7360
AdaBoost	0,7558	0,7503
XGBoost	0,7651	0,7690
GBoost	0,7721	0,7636

According to Table 8, Random Forest variants (RF5, RF10, RF15) did not outperform the best-performing single models. However, boosting ensemble architectures, specifically AdaBoost, XGBoost, and Gradient Boosting (GBoost), achieved noticeably higher F1 scores for the positive mortality class. The highest overall scores were obtained by GBoost on unnormalized data (0.7721) and XGBoost on normalized data (0.7690). Across all ensemble architectures, normalization again showed no substantial impact on overall predictive performance.

b. "three features" part

In this part, we omitted feature selection and only made the model selection for single models and ensemble models. All the models were trained with three features.

Table 9. **Performance of Single Models in "three features"**

Model	Unnormalized	Normalized
KNN3	0,5614	0,5644
KNN5	0,5558	0,5813
KNN7	0,5504	0,5764
NB	0,3815	0,3815
DT	0,5297	0,5392
LR	0,3775	0,3775
SVC(rbf)	/	0,5644
SVC(poly)	/	0,2556
SVC(sigmoid)	/	0,3789

Table 9 shows that results from normalized data that are slightly better than unnormalized ones, especially for KNN5, and KNN7. However, when comparing these results with those of the "all features" part, we can find that they are lower than

Table **7**.

This difference happened because models trained on "all features" had more patient data to learn from, helping them find clearer patterns. On the other hand, models using only "three features" had very limited information, making it harder for them to learn as effectively.

CONCLUSION

In conclusion, our study followed three straightforward steps to build and evaluate predictive models. First, we normalized all features to a uniform range between 0 and 1 to handle differences in

measurement scales. Second, we divided our experiments into two setups: a "three features" group matching the test dataset and an "all features" group exploring a broader set of biomarkers. Finally, we used cross-validation to select the best feature combinations, trained both single algorithms and ensemble models, and evaluated their performance on testing data.

Our results demonstrate that ensemble models consistently outperformed single models across both setups. Surprisingly, ensemble models using only three features achieved the highest overall performance, reaching a top F1 score of 0.8986 for predicting patient mortality, a result very close to the 0.90 benchmark reported by Yan et al. (2020). Normalization showed no noticeable impact on final accuracy. The strong performance of the three-feature models occurred because combining all top features at once in the full-feature setup introduced extra noise, while focusing on three key biomarkers gave the models a cleaner signal.

This study has two main limitations. First, we relied on standard ensemble models rather than building custom ensembles from our best individual models. Second, while filling missing values with 0 produced high scores, zero-imputation shifts the data distribution and can introduce prediction bias.

In future work, it is suggested to test different feature combinations in the full-feature setup rather than combining them all at once. It is also recommended to build custom ensembles by combining predictions from the best single models and explore safer imputation methods to eliminate bias while maintaining high predictive accuracy.

REFERENCES

Akter, T., Hossain, Md. F., Ullah, M. S., & Akter, R. (2024). Mortality Prediction in COVID-19 Using Time Series and Machine Learning Techniques. *Computational and Mathematical Methods*, *2024*(1), 5891177. https://doi.org/10.1155/2024/5891177

Gordan, M., Sabbagh-Yazdi, S.-R., Ismail, Z., Ghaedi, K., Carroll, P., McCrum, D., & Samali, B. (2022). State-of-the-art review on advancements of data mining in structural health monitoring. *Measurement*, *193*, 110939. https://doi.org/10.1016/j.measurement.2022.110939

Halwani, M. A., & Halwani, M. A. (2024). Prediction of COVID-19 Hospitalization and Mortality Using Artificial Intelligence. *Healthcare*, *12*(17), 1694. https://doi.org/10.3390/healthcare12171694

Hamar, Á., Mohammed, D., Váradi, A., Herczeg, R., Balázsfalvi, N., Fülesdi, B., László, I., Gömöri, L., Gergely, P. A., Kovacs, G. L., Jáksó, K., & Gombos, K. (2024). COVID-19 mortality prediction in Hungarian ICU settings implementing random forest algorithm. *Scientific Reports*, *14*(1), 11941. https://doi.org/10.1038/s41598-024-62791-9

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Khuluq, H., Astagiri Yusuf, P., Aryani Perwitasari, D., & Nguyen, T. (2024). Mortality prediction of COVID-19 patients using supervised machine learning. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *13*(4), 4472. https://doi.org/10.11591/ijai.v13.i4.pp4472-4479

Moulaei, K. (2021). Predicting Mortality of COVID-19 Patients based on Data Mining Techniques. *Journal of Biomedical Physics and Engineering*, *11*(5).

https://doi.org/10.31661/jbpe.v0i0.2104-1300

R. Han, Z. Liu, C. L. p. Chen, L. Xu, & G. Peng. (2020). Mortality prediction for COVID-19 patients via Broad Learning System. *2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, 837–842. https://doi.org/10.1109/ICCSS52145.2020.9336835

Sarwar, S., Shahzad, K., Fareed, Z., & Shahzad, U. (2021). A study on the effects of meteorological and climatic factors on the COVID-19 spread in Canada during 2020. *Journal of Environmental Health Science and Engineering*, *19*(2), 1513–1521. https://doi.org/10.1007/s40201-021-00707-9

Shanbehzadeh, M., Orooji, A., & Kazemi-Arpanahi, H. (2021). Comparing of data mining techniques for predicting in-hospital mortality among patients with covid-19. *Journal of Biostatistics and Epidemiology*, *7*(2), 154–169. Scopus.

Unpingco, J. (2016). *Python for Probability, Statistics, and Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-30717-6

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., … Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, *2*(5), 283–288. https://doi.org/10.1038/s42256-020-0180-7

Yildirim, S., Sunecli, O., & Kirakli, C. (2024). Mortality prediction with machine learning in COVID-19 patients. *Journal of Critical Care*, *81*, 154589. https://doi.org/10.1016/j.jcrc.2024.154589