COVID-19 Patient Outcome Prediction Using Selected Features from Emergency Department Data and Feed-Forward Neural Networks

Sophie Peacock[1], Mattia Cinelli[1], Frank S. Heldt[1(✉)], Lachlan McLachlan[1], Marcela P. Vizcaychipi[2,3], Alex McCarthy[2], Nadezda Lipunova[1], Robert A. Fletcher[1], Anne Hancock[1], Robert Dürichen[1], Fernando Andreotti[1], and Rabia T. Khan[1]

[1] Sensyne Health plc, Schrodinger Building, Heatley Road, Oxford Science Park, Oxford OX4 4GE, UK
stefan.heldt@sensynehealth.com
[2] Chelsea and Westminster Hospital NHS Foundation Trust, 369 Fulham Road, London SW10 9NH, UK
[3] Academic Department of Anaesthesia and Intensive Care Medicine, Imperial College London, Chelsea and Westminster Campus, 369 Fulham Road, London SW10 9NH, UK

Abstract. The severity of COVID-19 varies dramatically, ranging from asymptomatic infection to severe respiratory failure and death. Currently, few prognostic markers for disease outcomes exist, impairing patient triaging and treatment. Here, we train feed-forward neural networks on electronic health records of 819 confirmed SARS-CoV-2 positive patients admitted to a two-site NHS Trust hospital in London, England. To allow early risk assessment, the models ingest data collected in the emergency department (ED) to predict subsequent admission to intensive care, need for mechanical ventilation and in-hospital mortality. We apply univariate selection and recursive feature elimination to find the minimal subset of clinical variables needed for accurate prediction. Our models achieve AUC-ROC scores of 0.78 to 0.87, outperforming standard clinical risk scores. This accuracy is reached with as few as 13% of clinical variables routinely collected within the ED, which increases the practical applicability of such algorithms. Hence, state-of-the-art neural networks can predict severe COVID-19 accurately and early from a small subset of clinical variables.

Keywords: Machine learning · COVID-19 · Electronic health records

J. Ye et al. (Eds.): MobiHealth 2020, LNICST 362, pp. 323–335, 2021.
https://doi.org/10.1007/978-3-030-70569-5_21

1 Introduction

The novel severe acute respiratory syndrome virus 2 (SARS-CoV-2) has caused a pandemic outbreak of COVID-19 and a worldwide public health emergency. As of November 2020, the pandemic has led to more than 60 million confirmed cases and 1.5 million deaths [2]. While most COVID-19 patients have an asymptomatic infection or only suffer mild upper respiratory tract illness, the disease can progress to severe viral pneumonia with acute respiratory distress, respiratory failure and thromboembolic events that can lead to death [17,25,29]. Currently, few predictors for the transition to severe disease are known. However, an early identification of patients at-risk of severe outcomes may allow for faster intervention, improving treatment and therapy success.

The combination of state-of-the-art machine learning (ML) methods with electronic health records (EHRs) promises to predict patient deterioration with high precision [11,26]. Due to the scarcity of COVID-19 EHR data in the public domain, the majority of previous work has focused on statistical analyses or classical ML algorithms. Initial reports noted that factors such as age and underlying comorbidities can have an adverse effect on disease progression [7]. Zhou et al. used logistic regression on data of 191 COVID-19 positive patients to explore the risk factors for acute respiratory distress syndrome [31]. Similarly, Xie et al. applied logistic regression to the data of 299 COVID-19 positive patients to predict mortality [27]. Yan et al. [28] utilised XGBoost and EHR data of 375 COVID-19 patients in Wuhan, China to predict deterioration to a critical condition. While such studies provide insights into potential risk factors for severe COVID-19, most were conducted with limited patient numbers and data taken from both the patient's historical record and from throughout the current hospital admission [16,20,31]. The latter impairs an application to early patient triaging since, at the time of hospital presentation, the full EHR is rarely available. This problem is addressed by Jiang et al. [15] who applied ML methods to data available at the point of admission to hospital. However, with a sample size of just 53 patients the power of this study was limited.

Already prior to the COVID-19 pandemic, traditional risk scores were widely used in clinical practise to assess patient deterioration. Jones et al. explored the use of the sequential organ failure assessment (SOFA) score in combination with ML methods to forecast poor patient outcomes [16]. Using data collected from 248 patients over 2 years, they were able to predict in-hospital mortality by applying logistic regression to SOFA scores. Similarly, Scott et al. [21] have adopted the national early warning score (NEWS2) to predict the clinical outcome of patients. Yet, it remains unclear whether SOFA, NEWS2 or other similar clinical risk scores can be applied to COVID-19 patients.

A major obstacle to early patient triaging is the minimum number of clinical variables and, hence physiological tests, required to assess whether a patient is at risk. Feature selection methods, routinely applied in ML model development [4], can provide such a reduced feature set, retaining only the most informative clinical variables. Guyon and Elisseeff [12] introduce a number of methods which can be used to retain relevant information in a data set while reducing the number

of features. These methods can be split into filter, wrapper and embedded methods. Both filter and wrapper methods were previously used by Pourhomayoun and Shakibi [20] when predicting mortality in COVID-19 patients. In addition, Yan et al. used simple feature importance metrics to perform feature selection for predicting deterioration to a critical condition in COVID-19 patients [28].

We propose to use feed-forward neural networks to extract non-linear interactions between clinical variables and predict whether patients will deteriorate to severe COVID-19. We define deterioration to severe COVID-19 by three endpoints: admission to an adult intensive care unit (AICU), a need for mechanical ventilation, and in-hospital mortality. We perform feature selection to identify a minimal subset of clinical features that allows patient stratification and compare these subsets between endpoints. To facilitate early risk assessment, we focus our analysis on data available during a patient's emergency department (ED) visit at a hospital. Hence, the main contributions of this work are three–fold: 1. Early prediction of COVID-19 patients' risk to deteriorate to one of three clinical endpoints using neural networks; 2. evaluation of prediction performance over classical clinical risk scores; and 3. exploration of the minimal set of clinical features required for accurate patient stratification.

2 Methods

2.1 Data

Anonymised patient EHRs have been collected from a two-site NHS Trust hospital in London between January 1st and April 23rd 2020. All data were supplied according to internal information governance review, NHS Trust information governance approval, and General Data Protection Regulation (GDPR) procedures outlined under the Strategic Research Agreement (SRA) and relative Data Sharing Agreements (DSAs) signed by the NHS Trust and ourselves on 25th July 2018.

We analysed data from adult patients aged 18 to 100 and confirmed SARS-CoV-2 positive, as determined by quantitative reverse-transcription PCR (qRT-PCR). A total of 96 clinical features have been collected in the study, including patient demographics, vital signs, laboratory measurements and clinical observations. Of these 96 features, those with a coverage of at least 5% were retained. These 64 features are listed in the appendix in Table 3. Observations with multiple values were aggregated using the minimum, maximum, mean and last observation values to avoid biasing models on the number of test results. However, for blood test results typically only a single measurement is available within the ED stay of a patient, such that there is no distinction between the four aggregated values.

2.2 Cohort Definition

Study parameters included EHRs of 3229 patients. The data were filtered to include patients with confirmed SARS-CoV-2 infection (1158 patients), recorded

emergency department admission and subsequent ward stay, and their latest hospital admission being in 2020.

The patients were assigned in three cohorts (see Table 1): *Cohort A* was used to predict AICU admission. This cohort was divided into target patients who were admitted to an AICU at any time during their hospital stay, and control patients who were not. In the mechanical ventilation *Cohort B*, patients without clear information on oxygen supply were excluded; target patients required invasive mechanical ventilation, while control patients are those who required no or only minimal breathing assistance. For the mortality *Cohort C*, patients deceased during hospitalisation were considered target and discharged patients are included in the control group. Patients still hospitalised at the moment of study or deceased after hospitalisation were not considered.

Table 1. Patient numbers in study cohorts.

	Cohort A (AICU)	Cohort B (ventilation)	Cohort C (mortality)
Patients	819	818	508
Target	126 (15%)	62 (8%)	170 (33%)
Control	693 (85%)	756 (92%)	338 (67%)

2.3 Prediction Algorithms

EHR data from ED visits were used as inputs to a feed-forward neural network to predict patient outcomes. Hyper-parameter optimisation was carried out using Bayesian optimisation with Gaussian process as surrogate model using Keras Tuner [19]. Optimisation parameters included the number of fully connected hidden layers ($n_{layers} \in [0, 5]$) with ReLU activation functions containing a number of neurons per layer ($n_{neurons} \in [2, 96]$), before a single-neuron output layer with sigmoid activation. Batch normalisation with a batch size of six and dropout rate ($d \in [0, 0.5]$) were used after each hidden layer. Training used an Adam optimiser with binary cross-entropy loss and optimised learning rate ($n_{lr} \in [1e^{-4}, 1e^{-2}]$), for 100 epochs with early stopping. Optimisation was performed using the loss on the validation set from a nested stratified 80%/20% training/validation split derived from the training set of a 3-fold cross validation and the mean configuration was chosen (Table 4).

Prior to model training, features with less than 5% coverage were removed, missing values were imputed with a fixed value of -1 and the data were normalised using standard normalisation. Due to the large class imbalance, the minority class was oversampled using SMOTE [5].

The performance of ML algorithms was measured against the performance of the SOFA [18,23] and NEWS2 [1] scores, which are commonly used in clinical practice. The SOFA score was developed to evaluate morbidity in relation to organ dysfunction in critically ill patients [18,22]. Successive analyses have shown that SOFA scores are good indicators of prognosis [9,23]. The NEWS2 score aims to be a valid indicator of the patient's well-being at an early stage of their

hospitalisation. Less frequently it is used as predictor of patient outcome [8]. In our analysis we use the maximum SOFA and NEWS2 score for each patient while in the Emergency Department. Where data are missing, zero points are added to each score.

2.4 Model Validation

Stratified 3-fold cross validation was used in the training and evaluation of the neural network models. Performance of the models is reported as area under the curve (AUC) of the receiver operating characteristic (ROC). The variability across folds provides a measure of model stability. Since the SOFA and NEWS2 scores are deterministic, 3-fold cross validation was not carried out. For these models we measure performance by the AUC-ROC.

2.5 Feature Selection

Due to the large number of features, such as laboratory results, we expect a large amount of redundant information. We therefore apply feature selection methods to find the minimal subset of clinical features that reliably predicts each endpoint. This feature subset allows accurate predictions and easy application in real-world settings where data may be sparse. Dimensionality reduction techniques similar to Principal Components Analysis were not implemented. Although these methods would create a smaller set of features, measurements from all parameters would still be required and therefore these techniques are not beneficial in practice.

Two feature selection methods were considered [12]. First, we applied univariate selection, a filter method in which the number of features to keep is specified. The dependency between each feature and the target output was calculated using mutual information [3]. The most important features according to mutual information were retained. We also considered recursive feature elimination (RFE), a wrapper method which starts with all of the features and repeats a process of eliminating the least informative features until only a set number of features remains [13]. A neural network was trained on each data set and permutation feature importances were used to determine which features to discard in each round [10].

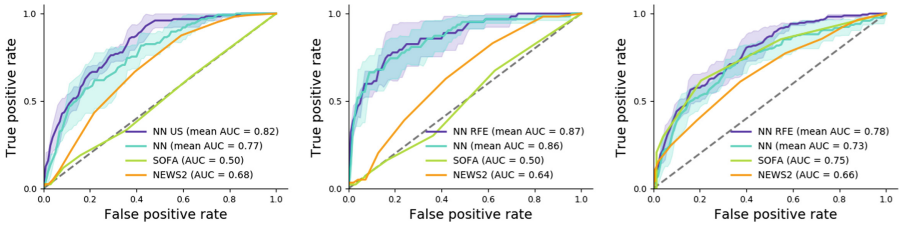
Feature selection was performed on the training set of each of the cross validation folds in order to obtain a feature list containing a specified number of features. For each feature selection method a grid-search was carried out to determine the optimal number of features to keep within the folds. The feature list was then used to train a model and make predictions. The feature lists presented in the appendix in Table 5 contain the union of the three feature lists obtained from each cross validation fold. Optimality was determined by AUC of ROC curve of the models obtained from the three cross validation folds for each endpoint.

3 Results

In the following, we first present baseline model performance when predicting three clinical endpoints for COVID-19 patients. The minimal subset of features for each endpoint is shown in the appendix (Table 5).

3.1 Neural Network Performance

Figure 1 and Table 2 show that the neural network with no feature selection outperforms both the SOFA and NEWS2 scores by a large margin when predicting AICU admission and mechanical ventilation. This is expected as the network is able to model complex relationships between multiple features and non-linear interactions. The difference in performance between our neural network and traditional scores is most pronounced when predicting a need for mechanical ventilation. In predicting mortality, the SOFA score performance shows a significant increase, while that of the neural network does not. Since the SOFA score was developed to evaluate morbidity this is to be expected [23]. The neural network model reaches an AUC of 0.73 when predicting mortality. While the model outperforms the NEWS2 score, it is not able to achieve a better result than prediction based on SOFA, which has an AUC of 0.75.



(a) AICU Admission (b) Mechanical Ventilation (c) Mortality

Fig. 1. Prediction performance for clinical endpoints. ROC curves of the neural network without (NN) and with feature elimination (NN RFE/NN US) and for SOFA and NEWS2. Solid lines and shaded areas indicate the mean and standard deviation across cross-validation folds, respectively. Dashed line indicates a random classifier.

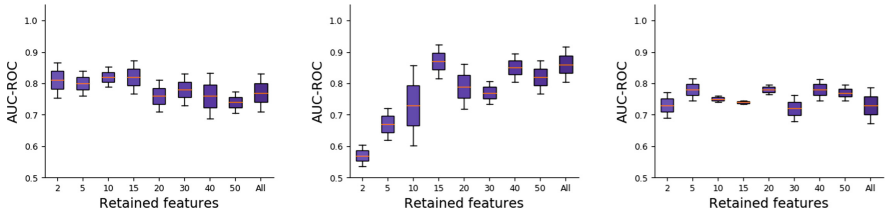
3.2 Performance with Feature Selection

Next, we use feature selection to identify the minimal subset of clinical variables required for accurate predictions. Overall, univariate feature selection performs best for predicting AICU admission, while RFE is best for predicting a need for mechanical ventilation and in-hospital mortality (see Table 2). Figure 2 shows the model performance over successively reduced feature sets, using univariate selection for AICU admission and RFE for the other two endpoints.

When predicting AICU admission the optimal number of features to keep is 10 in each cross validation fold (see Fig. 2a); the list of retained features across

Table 2. Predictive performance (AUC) for all endpoints. NN, neural network; US, univariate selection; RFE, recursive feature elimination. Standard deviation across cross validation folds is shown in brackets for NN models.

	Cohort A (AICU)	Cohort B (ventilation)	Cohort C (mortality)
SOFA	0.50	0.50	0.75
NEWS2	0.68	0.64	0.66
NN	0.77 (0.060)	0.86 (0.056)	0.73 (0.057)
NN + US	**0.82 (0.032)**	0.84 (0.047)	0.77 (0.035)
NN + RFE	0.78 (0.053)	**0.87 (0.054)**	**0.78 (0.035)**



(a) AICU Admission (b) Mechanical Ventilation (c) Mortality

Fig. 2. Prediction performance for feature sets of varying size. Boxes indicate AUC-ROC over cross-validation folds using univariate selection (a) and RFE (b, c), with the median marked by the orange lines and interquartile range by box edges.

all folds is included in the appendix in Table 5. By using univariate selection we achieve a significant increase in AUC of 5%.

As can be seen in Fig. 2b, for Cohort B the best performance is achieved by RFE using 15 features in each cross validation fold. The features used across all folds are listed in Table 5. This model achieves an AUC of 0.87, an improvement of just 1% over the neural network with no feature selection, potentially due to the baseline performance without feature selection already being high for this endpoint.

For prediction of in-hospital mortality, predictive performance of 0.78 AUC is attained using RFE with 5 features in each fold. The features used across all folds for predicting in-hospital mortality are shown in Table 5. The ROC curve for this model (Fig. 1c) shows an improvement in the AUC of 5%. This increase in performance allows the neural network with RFE to outperform the SOFA score by 3%.

The number of features in the optimal feature subset varies across the endpoints. When predicting AICU admission a large improvement in performance is gained by using a small feature list of just 10 features in each cross validation fold (25 unique features across all folds). Using 15 features per fold (36 overall) when predicting a need for mechanical ventilation leads to an improvement in predictive performance of just 1%. An improvement of 5% is also achieved through the retention of just 5 features per fold (10 overall) for predicting

in-hospital mortality. This improvement is especially significant as it enables the neural network model to outperform the SOFA score.

For all endpoints the overall feature list includes age and respiratory rate (Table 5). Markers of ethnicity are present for the prediction of a need for mechanical ventilation and in-hospital mortality. Vital sign measurements concerning temperature and fraction of inspired oxygen (FiO_2) are present for prediction of AICU admission and mechanical ventilation, while heart rate is retained for mechanical ventilation and in-hospital mortality. Although most of these features have very high coverage (both temperature and heart rate are above 99%), they are not consistently retained for all endpoints. A feature which has 100% coverage but is surprisingly discarded for prediction of AICU admission is sex.

As well as demographic and vital features, all overall feature lists include laboratory test results (Table 5). For the prediction of in-hospital mortality just 30% of features are laboratory tests, while for AICU admission and mechanical ventilation this figure is 72% and 61% respectively. We therefore see that the prediction of in-hospital mortality relies less on laboratory test results than the other two endpoints. While all overall feature lists contain a number of laboratory tests, there is a high degree of variability and only one test is present in all three feature lists - blood amylase. Various other laboratory tests are present for prediction of both AICU admission and mechanical ventilation; bicarbonate, creatinine, blood lactate, oxygen partial pressure, blood potassium and different forms of haemoglobin. Some of these laboratory tests, such as bicarbonate, oxygen partial pressure and blood lactate, have coverage of around 27%, but are retained over tests such as blood white cells or blood monocyte count, which have coverage of 84% but are not included in any of the three overall feature lists.

4 Discussion

This work was motivated by the need to predict whether patients deteriorate to severe COVID-19 early during their hospital stay and to provide clinicians with a minimal subset of clinical features which allow risk prediction. To address these points, we trained neural network models which use EHR data from COVID-19 patients' ED admissions to predict one of three endpoints: admission to AICU, need for mechanical ventilation, and in-hospital mortality. We have shown that feed-forward neural networks can achieve better predictive performance on the first two endpoints than traditional risk scores. Neural networks without feature selection were not able to outperform the SOFA score for predicting in-hospital mortality, possibly due to the SOFA score being developed to predict morbidity. Implementing feature selection using univariate selection and RFE enabled us to identify the minimal subset of clinical features required for early risk assessment of COVID-19 patients. For AICU admission, need for mechanical ventilation and in-hospital mortality, performance was improved by 5%, 1% and 5% respectively. For predicting in-hospital mortality, feature selection allowed us to achieve a predictive performance 3% higher than that of the SOFA score.

Aside from an improvement in predictive performance, a model requiring fewer features is extremely beneficial in its applicability to real-world scenarios. A significantly reduced set of required inputs means that the model can be applied in settings where data may be sparse and not all of the original features are available. Having to collect fewer data points in order to make a prediction increases the accessibility of the model and allows clinicians to prioritise testing.

The predictive performance achieved by our neural network models is comparable to previous work using an XGBoost model [14]. Feature selection methods are employed by Pourhomayoun and Shakibi [20] to reduce their feature set from 112 to 42 features, although they do not also present results for models trained using the entire feature set. Our findings that age and indicators of oxygenation status always remain in the final feature set are consistent with this work. Age in particular is consistently found to be an important feature in previous works [27,31]. Conversely, our finding that sex is not retained for prediction of AICU admission differs from previous works [24,30].

While laboratory test results are included in the feature list for all endpoints, there is not a large degree of consensus regarding which tests are most informative. Our finding that features relating to haemoglobin are retained for two out of three endpoints are consistent with those of Jiang et al. [15]. A surprising finding of this work which may invite further analysis is the absence of heart rate in the overall feature list for prediction of AICU admission, and of FiO_2 and temperature for predicting in-hospital mortality. Temperature in particular is a common indicator for severe viral infection [6].

Taken together, our analysis and previous studies suggest that patient age, demographic information and measures of oxygenation status, such as respiratory rate and FiO_2 level, are primary indicators of poor outcomes in COVID-19 patients. Prioritising the measurement and clinical assessment of these variables may improve early patient triaging.

This work uses EHR data captured during a patient's ED visit. While this more accurately reflects the data available in practice, it may well limit the performance of our models. Augmenting the data set with patients' medical history may be beneficial, particularly in predicting mortality where a patient's chance of survival may be heavily influenced by their comorbidities and other medical history. While our data set is comparatively large in relation to previous COVID-19 studies [27,29], further improvements could be made with access to more data. Longitudinal data from other hospitals in different locations could improve the generalisability of our models. A significantly larger data set would also make it feasible to train more complex, deeper neural networks which may achieve higher prediction performance. One future approach to overcome data availability issues is the use of transfer learning on other respiratory diseases or multi-task learning on several clinical endpoints simultaneously.

In conclusion, our models show that state-of-the-art neural networks can predict severe COVID-19 accurately from sparse, clinical data. Importantly, we are able to produce a minimal subset of clinical variables required for early risk assessment of COVID-19 patients. Models trained on this minimal subset of features can be used by clinicians with limited data available to them to stratify patients into risk groups.

Acknowledgments. This work uses data collected from patients by the NHS as part of their care and support. Thus, we would like to thank all those involved for their contribution. Special thanks are due to the Chelsea and Westminster NHS Foundation Trust COVID-19 AICU Consortium, comprising all critical care personnel who were part of the delivery of care during the COVID-19 pandemic.

A Clinical Features

Table 3 contains all clinical features with over 5% coverage.

Table 3. All clinical features with at least 5% coverage

Age	Blood Glucose
Ethnicity	Blood Haematocrit
Sex	Blood Haemoglobin
FiO_2 level POC	Blood Lactate
Heart Rate	Blood Lactate Dehydrogenase Level
Respiratory Rate	Blood Lymphocyte Count
Temperature	Blood Magnesium
Blood Activated Partial Thromboplastin Time	Blood Mean Corpuscular Haemoglobin Concentration
Blood Adjusted Calcium	Blood Mean Corpuscular Haemoglobin
Blood Alanine Aminotransferase	Blood Mean Corpuscular Volume
Blood Albumin	Blood Mean Platelet Volume
Blood Alkaline Phosphatase	Blood Methaemoglobin
Blood Amylase	Blood Monocyte Count
Blood Anion Gap	Blood Neutrophil Count
Blood Base Excess	Blood Nucleated Red Blood Cell Count
Blood Basophil Count	Blood Oxygen PO_2 Partial Pressure
Blood Bicarbonate	Blood Oxyhaemoglobin
Blood Bilirubin Total	Blood pH
Blood C Reactive Protein	Blood Phosphate
Blood Calcium	Blood Platelet Count
Blood Calcium Ionised	Blood Potassium
Blood Carboxyhaemoglobin	Blood Prothrombin Time
Blood Chloride	Blood Red Blood Cell Count
Blood Cortisol	Blood Red Cell Distribution Width
Blood Creatine Kinase	Blood Sodium
Blood Creatinine	Blood Thyroid Stimulating Hormone
Blood D Dimer	Blood Thyroxine T4
Blood Deoxyhaemoglobin	Blood Total Protein
Blood Eosinophil Count	Blood Troponin T
Blood Ferritin	Blood Urea
Blood Fibrinogen	Blood White Cells
Blood Globulin	Brain Natriuretic Peptide

B Model Hyper-parameters

Table 4 contains the optimal model hyper-parameters for each endpoint.

Table 4. Optimal model hyper-parameters for each endpoint.

	Cohort A (AICU)	Cohort B (ventilation)	Cohort C (mortality)
Hidden layers	2	3	2
Neurons per layer	31	35	28
Dropout rate	0.12	0.30	0.26
Learning rate	0.002	0.005	0.002

C Feature Lists

Table 5 contains the features retained in the final trained models for predicting each endpoint. This list is the union of the features retained over the three cross validation folds for each endpoint.

Table 5. Overall features retained for each endpoint

AICU admission	Mechanical ventilation	Mortality
Age	Age	Age
Last alanine aminotransferase	Eth black african	Eth asian indian
Last amylase	Eth black caribbean	Eth asian pakistani
Last bicarbonate	Eth other chinese	Eth black other
Last blood ldh level	Eth unknown	Max amylase
Last nucleated red blood cell count	Eth white other	Max heart rate
Last oxyhaemoglobin	Last amylase	Mean blood ferritin
Last respiratory rate	Last blood lactate	Mean respiratory rate
Max anion gap	Last blood potassium	Min blood bilirubin total
Max blood ldh level	Last blood mean corpuscular haemoglobin mch	Sex
Max blood phosphate	Last deoxyhaemoglobin	
Max creatinine	Last FiO_2 level	
Max FiO_2 level	Last haemoglobin	
Max oxygen partial pressure	Last MCHC	
Max red blood cell width	Last mean platelet volume	
Max respiratory rate	Last respiratory rate	
Max temperature	Max amylase	
Mean alanine aminotransferase	Max carboxyhaemoglobin	
Mean blood lactate	Max FiO_2 level	
Mean blood ldh level	Max mean platelet volume	
Mean blood potassium	Max respiratory rate	
Mean fibrinogen	Max temperature	
Mean FiO_2 level	Mean blood magnesium	
Mean respiratory rate	Mean blood urea	
Min blood ldh level	Mean blood total protein	
	Mean FiO_2 level	
	Mean lymphocyte count	
	Mean MCHC	
	Min bicarbonate	
	Min creatinine	
	Min deoxyhaemoglobin	
	Min haemoglobin	
	Min heart rate	
	Min mean platelet volume	
	Min oxygen partial pressure	
	Sex	

References

1. National early warning score (news) 2, April 2020. https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2
2. Situation update worldwide, as of 16 July 2020 (2020). https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. Accessed 16 July 2020
3. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**, 245–271 (1997)
4. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
6. Chen, J., et al.: Clinical progression of patients with COVID-19 in Shanghai, China. J. Infect. **80**(5), e1–e6 (2020)
7. Chen, T., et al.: Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. BMJ **368**, m1091 (2020)
8. Churpek, M.M., et al.: Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit. Am. J. Respir. Crit. Care Med. **195**(7), 906–911 (2017)
9. Ferreira, F.L., Bota, D.P., Bross, A., Mélot, C., Vincent, J.L.: Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA **286**(14), 1754–1758 (2001)
10. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**(177), 1–81 (2019)
11. Goldstein, B.A., Navar, A.M., Pencina, M.J., Ioannidis, J.: Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J. Am. Med. Inform. Assoc. **24**(1), 198–208 (2017)
12. Guyon, I.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
13. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002). https://doi.org/10.1023/A:1012487302797
14. Heldt, F.S., et al.: Early risk assessment for COVID-19 patients from emergency department data using machine learning. medRxiv (2020)
15. Jiang, X., et al.: Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Comput. Mater. Continua **63**(1), 537–551 (2020)
16. Jones, A.E., Trzeciak, S., Kline, J.A.: The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. Crit. Care Med. **37**(5), 1649–1654 (2009)
17. Klok, F.A., et al.: Incidence of thrombotic complications in critically ill ICU patients with COVID-19. Thromb. Res. **191**, 145–147 (2020)
18. Lambden, S., Laterre, P.F., Levy, M.M., Francois, B.: The SOFA score-development, utility and challenges of accurate assessment in clinical trials. Crit. Care **23**(1) (2019). Article number: 374
19. O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al.: Keras Tuner (2019). https://github.com/keras-team/keras-tuner

20. Pourhomayoun, M., Shakibi, M.: Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. medRxiv (2020)
21. Scott, L.J., Redmond, N.M., Tavaré, A., Little, H., Srivastava, S., Pullyblank, A.: Association between national early warning scores in primary care and clinical outcomes: an observational study in UK primary and secondary care. Br. J. Gen. Pract. **70**(695), e374–e380 (2020)
22. Soo, A., et al.: Describing organ dysfunction in the intensive care unit: a cohort study of 20,000 patients. Crit. Care **23**(1), 186 (2019)
23. Vincent, J.L., et al.: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med. **22**(7), 707–710 (1996). On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine
24. Wang, D., et al.: Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA **323**(11), 1061–1069 (2020)
25. Wu, Z., McGoogan, J.M.: Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. JAMA **323**(13), 1239–1242 (2020)
26. Wynants, L., et al.: Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. BMJ **369**, m1328 (2020)
27. Xie, J., et al.: Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19 (2020)
28. Yan, L., et al.: Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. MedRxiv (2020)
29. Yang, X., et al.: Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir. Med. **8**(5), 475–481 (2020)
30. Zheng, Z., et al.: Risk factors of critical & mortal COVID-19 cases: a systematic literature review and meta-analysis. J. Infect. **81**(2), e16–e25 (2020)
31. Zhou, F., et al.: Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet **395**, 1054–1062 (2020)