Machine learning discovery of distinguishing laboratory features for severity classification of COVID-19 patients

Yang Xiao1 | Li Yan2 | Mingyang Zhang1 | Kent E. Pinkerton3 | Haosen Cao1
Ying Xiao1 | Wei Li4 | Shuai Li1 | Yancheng Wang1 | Shusheng Li2
Zhiguo Cao1 | Gary Wing-Kin Wong5 | Hui Xu6 | Hai-Tao Zhang1

1School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
2Department of Emergency, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
3Department of Pediatrics, School of Medicine, Department of Anatomy, Physiology and Cell Biology, School of Medicine, School of Veterinary Medicine, University of California, Davis, CA, USA
4Biomedical Engineering Institute, School of Medicine, Department of Anatomy, Physiology and Cell Biology, School of Medicine, School of Veterinary Medicine, University of California, Davis, CA, USA
5Department of Anesthesiology, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Abstract
The exponential spread of COVID-19 worldwide is evident, with devastating outbreaks primarily in the United States, Spain, Italy, the United Kingdom, France, Germany, Turkey and Russia. As of 1 May 2020, a total of 3,308,386 confirmed cases have been reported worldwide, with an accumulative mortality of 233,093. Due to the complexity and uncertainty of the pathology of COVID-19, it is not easy for front-line doctors to categorize severity levels of clinical COVID-19 that are general and severe/critical cases, with consistency. The more than 300 laboratory features, coupled with underlying disease, all combine to complicate proper and rapid patient diagnosis. However, such screening is necessary for early triage, diagnosis, assignment of appropriate level of care facility, and institution of timely intervention. A machine learning analysis was carried out with confirmed COVID-19 patient data from 10 January to 18 February 2020, who were admitted to Tongji Hospital, in Wuhan, China. A softmax neural network-based machine learning model was established to categorise patient severity levels. According to the analysis of 2662 cases using clinical and laboratory data, the present model can be used to reveal the top 30 of more than 300 laboratory features, yielding 86.30% blind test accuracy, 0.8195 F1-score, and 100% consistency using a two-way patient classification of severe/critical to general. For severe/critical cases, F1-score is 0.9081 (i.e. recall is 0.9050, and precision is 0.9113). This model for classification can be accomplished at a mini-second-level computational cost (in contrast to minute-level manual). Based on available COVID-19 patient diagnosis and therapy, an artificial intelligence model paradigm can help doctors quickly classify patients with a high degree of accuracy and 100% consistency to significantly improve diagnostic and classification efficiency. The discovered top 30 laboratory features can be used for greater differentiation to serve as an essential supplement to current guidelines, thus creating a more

Correspondence
Hai-Tao Zhang, School of Artificial Intelligence and Automation, Key Laboratory of Image Processing and Intelligent Control, The State Key Lab of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, China. Email: zht@mail.hust.edu.cn
Hui Xu, Department of Anesthesiology, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430074, China. Email: huxi@tjh.tjmu.edu.cn
Gary Wing-Kin Wong, Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong, China. Email: wingkinwong@cuhk.edu.hk
Zhiguo Cao, National Key Laboratory of Science and Technology on Multi-spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China. Email: zgcao@hust.edu.cn

Funding information
The COVID-19 Prompt Response Research Special Project from Huazhong University of Science and Technology, Grant/Award Numbers: 2020kfXYJ113, 2020kfXYJ023; The special

Yang Xiao, Li Yan, Mingyang Zhang and Kent E. Pinkerton contributed equally to this study.
INTRODUCTION

The exponential spread of Coronavirus disease 2019 (COVID-19) infection worldwide has caused great concern [1–6] Guidelines provided by the National Health Commission of China have categorised COVID-19 patients into four classes that is 1) mild, 2) moderate, 3) severe, and 4) critical. While these guidelines focus on indicators of the respiratory system, overall health conditions should also be considered for complete diagnostic purposes. Due to the complexity, latency, and uncertainty of COVID-19 pathology, these four severity classes are not consistently easy to distinguish. Other factors, such as age, organ status, underlying diseases, and other complications, can also be classified along with more than 300 clinical and laboratory features to further complicate the classification of COVID-19 disease status. The purpose of this study is to prompt a screening of the severe and critically ill cases to facilitate early triage according to laboratory data obtained at admission. It has been claimed that ‘the proportion of mild and asymptomatic cases versus severe and fatal cases is currently unknown for 2019-nCoV which is a knowledge gap that hampers realistic assessment of the virus’s epidemic potential and complicates the outbreak response’ [7]. From the public health stand point, initial severity screening thus becomes an indispensable step. That is, the mild/moderate cases could benefit from home isolation or tele-medical care, not strict hospitalisation, whereas the severe/critical patients will need early transfer to secondary/tertiary care.

Nonlinear projection from a low-to-high-dimensional feature demonstrates extremely mixed features, thus complicating the distinction between the two levels of infection/disease severity for 1) mild/moderate and 2) severe/critical classifications. Due to these challenges, we sought to extract specific, distinct indicators to assess the severity of COVID-19 in patients, based on more than 300 laboratory features using artificial intelligence (AI) or machine learning (ML) technology, to assist physicians and health providers to more precisely differentiate the severity classification of patients in a time-efficient and consistent manner. Using the data from 2662 COVID-19 cases confirmed by three experienced physicians (i.e. two deputy-chiefs and one physician-in-charge, physicians with 22, 20, and 11 years of clinical experience, respectively) jointly according to established guidelines [8], a neural network (NN)-based ML method was developed. The motivation for why choosing NN as the classifier is due to the key insight that, the research task of severity classification of COVID-19 patients is actually a fine-grained non-linear pattern recognition problem and multi-layer NN can provide strong non-linear pattern fitting capacity [9]. Using this technology, the top-30 features were revealed, together with age, to rapidly distinguish the most severe cases from those with lesser (milder) symptoms. This methodology resulted in 86.30% blind test accuracy, 0.8195 F1-score, and 100% consistency with a diagnostic time reduction going from minutes to seconds.

This AI technology has the capacity to screening out severe and critical patients with a high risk of death from other patients, to provide assistance in the identification of those patients in greatest need of medical intervention, including advanced ventilation therapy, to circumvent irreversible pathological changes. Prompt, consistent and accurate classification to facilitate early triage, diagnosis, and therapy is of utmost importance. Moreover, this technology revealed the top 30 laboratory features out of more than 300 for definitive differentiation of clinical information critical to current guidelines.

RELATED WORK

COVID-19 was first described in late December in Wuhan, China. The disease has spread quickly to almost all over the world. The treatment is primarily supportive as there is no proven specific treatment for this infection. Over 80% of patients develop mild disease and will recover uneventfully while the others will develop more severe disease. Although the main public health measure to control the community outbreak is to identify and isolate the infected patients early on so as to stop the chain of transmission, approximately 5% to 8% of patients may require intensive care support. Early identification of this group of patients is important such that they can be managed at appropriate health care facilities to institute necessary treatments in a timely manner. We searched PubMed on 19 September 2020, for studies using keywords including ‘COVID-19’, ‘SARS-CoV-2’, ‘classification’, ‘AI’, and ‘ML’. There were a total of 172 publications and 107 of them were original studies. Among these 107 original studies, 27 of them described using AI or ML for the analyses of radiological images [10–12] to improve the diagnostic accuracy of COVID-19. There was only one published study [13] using AI or ML to evaluate clinical and laboratory features for the classification of severity of COVID-19. The main differences between our study and it lie in three folders. First, the number of the patients involved in this previous work is only 137, while we have 2662 patients to conduct much larger-scale investigation to verify the generality of the proposed COVID-19 patient severity-level classification approach. Secondly, this previous study uses SVM [14] for severity-level categorisation, whereas it is revealed that this mission is a fine-grained non-linear pattern recognition problem [15] and multi-layer neural network is applied to seek stronger non-linear pattern fitting capacity [9]. Last, during the phase of feature selection, we more focus on the discriminative power of laboratory indicators rather than inter-feature redundancy [13].
METHODS

Data collection

For this retrospective, single-centre study, we collected the electronic records of 2662 validated COVID-19 patients from 10 January to 18 February 2020 admitted to Tongji Hospital, in Wuhan, China. We derived epidemiological, demographic, clinical, laboratory, drugs, nursing, and outcome data from the electronic medical records of all patients. For all 2662 patients, data were collected on admission to the Tongji Hospital, in Wuhan, China.

Cases with at least one of the following were identified as being COVID-19 positive (infected): 1) COVID-19 nucleic acid present in either respiratory or blood samples by RT-PCR; or 2) viral genetic sequence detected in respiratory or blood samples with high homology to COVID-19.

The COVID-19 severity level for all 2662 patients was jointly determined by three physicians, including two deputy-chief and one physician-in-charge, physicians with 22, 20, and 11 years of clinical experience, respectively. The four severity levels defined below are in accordance with the established guidelines [8].

1. Mild: clinical symptoms are mild, no signs of the pneumonia found on imaging;
2. Moderate: fever, respiratory tract, and other symptoms, evidence of pneumonia by imaging;
3. Severe: \( RR \geq 30 \text{ bpm or } \text{SPO}_2 \leq 93\% \text{ at rest, or } \text{PaO}_2/\text{FiO}_2 \leq 300 \text{ mmHg}; \)
4. Critical: shock, mechanical ventilation required, or ICU admission due to MODS.

Statistical analysis

All data were confirmed using SPSS 26.0. The continuity variables of normal and non-normal distribution were described by mean standard \( \pm \) deviation and median/quartile, respectively. The general data were tested for normality. The Kolmogorov–Smirnov (K–S) test was used to examine whether the single sample was from a particular distribution, and a single sample K-S test was later adopted to test the normality of the general data. In essence, results with \( \alpha = 0.05 \) and \( P < 0.05 \) indicates that the sample does not fit a normal distribution. Therefore, in Table 1, mean \( \pm \) standard deviation was used to describe age, which conforms to a normal distribution, whereas median was applied to represent other continuous variables, which showed non-normal distribution.

NN-based learning

To achieve fine-grained severity classification, a discriminative laboratory feature must meet the following two requirements: 1) different classes are to be sufficiently distinctive, and 2) the intra-class deviation should be sufficiently small to avoid class confusion. In other words, for certain feature \( f_i \), \( j = 1, \ldots, N_F \), where \( N_F \) is the feature number, the inter-class deviation [15]

\[
\text{Dev}_{ij}^{\text{inter}} = \sqrt{\frac{\sum_{k=1}^{N_F} (f_{ik} - \bar{f}_j)^2}{N_k - 1}},
\]

is calculated as \( \text{Dev}_{ij}^{\text{inter}} = \sqrt{\sum_{k=1}^{N_F} (f_{ik} - \bar{f}_j)^2}/N_k - 1 \), where \( N_k = 2 \) is the severity level number, \( \bar{f}_j \) denotes the mean value of \( f_i \) over all the severity levels and \( \bar{f}_j \) denotes the mean value of \( f_i \) on the \( j \)th severity level, \( j = 1, \ldots, N_L \). In contrast, the intra-class deviation [15] \( \text{Dev}_{ij}^{\text{intra}} \) is given by \( \text{Dev}_{ij}^{\text{intra}} = \sqrt{\sum_{k=1}^{N_F} (f_{ik} - \bar{f}_j)^2}/N_k - 1 \), where \( \bar{f}_k \) and \( N_j \) denote the \( k \)th sample from and the sample number of the \( j \)th severity level, respectively. Thereby, the overall discriminative index of \( f_i \) is calculated as

\[
\text{Dis}_j = \frac{\text{Dev}_{ij}^{\text{inter}}}{\text{Dev}_{ij}^{\text{intra}}}. \tag{1}
\]

Evidently, as shown in Figure 1(b) and (e), the higher the index \( \text{Dis}_i \), the more discriminative identification is achieved.

In addition to the revealed top-30 laboratory features, age is considered an indispensable non-laboratory feature, adopted as a supplementary feature due to its high correlation to COVID-19 pathology [3]. Due to the different value ranges, the selected 30 + 1 features are first normalised by \( f_i^{\prime} = f_i - \bar{f}_i/\text{Dev}_i \), before being introduced into the input layer of the NN, where \( \text{Dev}_i \) denotes the standard deviation of feature \( f_i \) disregarding the class information. Meanwhile, ‘0’ padding is also used on input feature. A three hidden-layer NN is established in Figure 2(a). Therefore, the input layer consists of 31 input units corresponding to the 30 + 1 features, while the output layer contains two output units representing the two severity levels, and the three hidden layers consist of 64, 128, and 64 neurons for non-linear feature mapping via the leaky ReLU activation function [16]. The main intuitions for choosing NN for COVID-19 patient severity-level categorisation rather than the other classifiers lie in the following three facts: (1) COVID-19 patient severity-level categorisation is actually a non-linear fine-grained pattern categorisation problem, as revealed in Figure 2 (b) and (c) (2) multi-layer NN (e.g., NN-based deep learning); is of strong non-linear fitting capacity when training samples are sufficient to avoid overfitting [9]; and (3) till 23 September 2020, there have been 31,375,325 confirmed cases of COVID-19 globally according to WHO Corona virus disease (COVID-19) Dashboard (i.e., https://covid19.who.int/), thus acquiring sufficient COVID-19 patient samples is no longer a bottleneck to ensure training performance.

Due to the fine-grained categorisation challenge of COVID-19 severity classification and the overfitting risk of NN, A-softmax loss function [17] is used to drive the training procedure of NN. As shown in Figure 1(a), an angle-based decision boundary is established with a sufficiently extended angular margin among the pair-wise classes, so as to stretch the inter-class distance in the discriminative index of Equation (1). The non-linear evolution procedure of the input COVID-19 laboratory features distributed through the present
| Characteristics | Overall |
|-----------------|---------|
| Age, mean (SD), y | 58.69 (30.52) |
| **Gender** | n (%) |
| Male | 1343 (50.47) |
| Female | 1318 (49.53) |
| **Symptoms on onset** | n (%) |
| Fever | 1054 (39.61) |
| Cough | 400 (15.03) |
| Fatigue | 41 (1.54) |
| Dyspnea | 49 (1.84) |
| Chest distress | 43 (1.62) |
| Muscular soreness | 10 (0.38) |
| Diarrhoea | 43 (1.62) |
| Sore throat | 27 (1.01) |
| Anorexia | 14 (0.53) |
| **Lab test** | |
| High-sensitivity C-reactive protein, median (range, Q1-Q3), (mg/L) | 30 (6.4–74.5) |
| Lactate dehydrogenase, median(range, Q1–Q3), U/L | 285 (223–384) |
| Albumin, median (range, Q1-Q3), g/L | 34.4 (31.2–37.8) |
| Free T3, median (range, Q1-Q3), pg/mL | 3.63 (3.03–4.32) |
| Cholinesterase, median (range, Q1-Q3), U/L | 6160 (4941–7431) |
| Interleukin 2 receptor, median (range, Q1-Q3), U/mL | 701 (478–1053) |
| Prealbumin, median (range, Q1-Q3), g/L | 138 (80–208) |
| Neutrophils, median (range, Q1-Q3), 10⁹/L | 3.9 (2.69–5.68) |
| Myoglobin, median (range, Q1-Q3), mg/mL | 45 (28.3–91.6) |
| Aspartate transaminase, median (range, Q1-Q3), U/L | 27 (19–40) |
| Globulin, median (range, Q1-Q3), g/L | 33.3 (30.0–36.9) |
| Oxygen saturation, median (range, Q1-Q3), % | 97.9 (95–99.1) |
| D-D Dimer quantification, median (range, Q1-Q3), µg/mL | 0.91 (0.45–2.04) |
| Fibrinogen, median (range, Q1-Q3), g/L | 4.95 (3.9–5.97) |
| Fibrin degradation products, median (range, Q1-Q3), mg/L | 4.0 (4–8.2) |
| Erythrocyte sedimentation rate, median (range, Q1-Q3), mm/h | 36 (19–62) |
| Ferritin, median (range, Q1-Q3), µg/L | 640.8 (366.8–1207.3) |
| Tumour necrosis factor α, median (range, Q1-Q3), pg/mL | 8.3 (6.3–11.2) |
| Prothrombin activity, median (range, Q1-Q3), % | 91 (82–99) |
| Alanine transaminase, median (range, Q1-Q3), U/L | 23 (15–38) |
| eGFR, median (range, Q1-Q3), mL/min/1.73 m² | 91 (76.5–102) |
| Leucocyte count, median (range, Q1-Q3)×10⁹/L | 5.68 (4.37–7.52) |
| Urea, median (range, Q1-Q3), mmol/L | 4.4 (3.4–6.1) |
| Cystatin C, median (range, Q1-Q3), mg/L | 1.03 (0.87–1.24) |
| Calcium, median (range, Q1-Q3), mmol/L | 2.12 (2.05–2.2) |
| Sodium, median (range, Q1-Q3), mmol/L | 139.3 (136.7–141.6) |
TABLE 1 (Continued)

| Characteristics                                      | Overall |
|------------------------------------------------------|---------|
| High density lipoprotein, median (range, Q1-Q3), mmol/L | 0.89 (0.75-1.06) |
| γ-glutamyl transpeptidase, median (range, Q1-Q3), U/L   | 29 (19-52) |
| Bicarbonate root, median (range, Q1-Q3), mmol/L        | 23.8 (21.9-25.5) |
| pCO₂, median (range, Q1-Q3), mmHg                      | 37.5 (32.9-41.5) |

*The listed 30 features are selected from over 300 ones via machine learning discovery.

NN model is shown in Figure 3(d)–(g), which corresponds to the training set from 1 running around randomly picked from the 5-fold cross validation [15]. Apparently, with the assistance of the delicately designed lost function [17] in Figure 1(a), dramatic non-linear feature warping happens during the evolution, and hence the two severity classes become more and more distinct through the three neural layers. To this end, the present NN model yields consistent classification from what appears to be seemingly linearly inextricable of the two kinds of samples.

In the experiments, Diederik and Adam [18] is used as the optimiser for NN training of batch size 32. The learning rate was set to 0.001 with a decay rate of 0.9 per 25 training epochs [18]. A total of 60 epochs were used. Weight decay was set to 0.001. A 5-fold cross-validation [15] protocol was used for the NN test.

RESULTS

The clinical/laboratory features and treatment of 2662 COVID-19 patients admitted to Tongji Hospital are shown in Table 1. There are three available classes of severity jointly labelled by the three involved experienced physicians (i.e. two deputy-chief and one physician-in-charge, physicians with 22, 20, and 11 years of clinical experience) via majority voting towards all the 2662 COVID-19 patients that are moderate, severe, and critical cases, according to the guidelines provided by the National Health Commission of China and the practical clinical situation during the data collection period. Since the latter two classes are more essential to identify for early diagnosis and advanced therapy, we merged them into a severe + critical (in abbr. SC) class. Meanwhile, moderate was categorised into the general class that includes mild and moderate to facilitate the general and SC case categorisation. The intuition is that the general cases could benefit from home isolation or tele-medical care, not strict hospitalisation, whereas the SC patients need early transfer to secondary/tertiary care. Using this classification scheme, we have established a NN learning-based severity classification model which determined the top 30 laboratory features along with one non-laboratory feature (i.e., age) so as to yield the following general VS SC classification results.

For integrity, we have implemented data filtering to remove 523 samples with incomplete laboratory features from the 2662 cases. Particularly, for one sample if the available feature number among the selected 30 + 1 features is less than 15, it will be discarded, to alleviate the incomplete information problem within our retrospective study. Among the remaining 2139 samples, the numbers of general and SC cases are 539 and 1600 respectively to conduct the binary severity categorisation test. ‘0’ padding on the missing features is also executed to them. In addition, a 5-fold cross validation [15] is subsequently implemented on the remaining 2139 integral samples. The metrics of Accuracy [19], Precision [19], Recall [19], and
F1-score [19] are calculated for performance evaluation over all the five validation sets jointly.

Logistic regression [20], Gaussian Bayes classifier [21], K-nearest neighbours (KNN) classifier [22], linear SVM [14], non-linear Polynomial SVM, non-linear Sigmoid SVM [14], and quadratic discriminant analysis (QDA) classifier [23] are used to conduct the performance comparison with the proposed A-softmax NN model, with the same input features.

**Top-30 discovered laboratory features**

As a well-accepted ML clustering index, the discriminative criteria is selected as a measure $\text{Dis}_i$ ($i$ is the feature sequential number) calculated by inter-class distance over intra-class distance given in Equation (1). Accordingly, the top-30 laboratory features ($i = 1, \ldots, 30$) with the highest $\text{Dis}_i$ are screened out for COVID-19 severity differentiation, which are
Figure 3 The $D_{ij}$ ranking (a) and the radar map (b) of the top-30 laboratory features with feature sequential number $i$. For each feature, the colour bar reveals the class-wise mean value and standard deviation, with feature normalisation in advance. The blue curve indicates the feature-wise distribution of $D_{ij}$. The radar map in (b) is exhibited after normalisation with details given in Method. Here, ‘severe + critical’ is abbreviated to SC in the main body.
listed sequentially as below: #1 high-sensitivity C-reactive protein (Hs-CRP), #2 lactate dehydrogenase (LDH), #3 Albumin, #4 Free T3, #5 Cholinesterase, #6 Interleukin 2 receptor, #7 Prealbumin, #8 Neutrophils, #9 Myoglobin, #10 Aspartate transaminase, #11 Globulin, #12 Oxygen saturation, #13 D-D dimer quantification, #14 Fibrinogen, #15 Immature reticulocyte, #16 Fibrin degradation products, #17 Erythrocyte sedimentation rate, #18 Ferritin, #19 Tumour necrosis factor α, #20 Prothrombin activity, #21 Alanine transaminase, #22 eGFR, #23 Leucocyte count, #24 Urea, #25 Cystatin C, #26 Calcium, #27 Sodium, #28 High density lipoprotein, #29 γ-glutamyl transpeptidase, #30 Bicarbonate root as shown in FIG. 3(a) and 3(b). Therein, the evolution of the index $D_{ij}$ is exhibited by the blue curve along with the above-mentioned sequential features, which drops quickly showing their discriminative capabilities.

**Fine-grained identification of COVID-19 severity**

To visualise the challenge of the COVID-19 severity classification, we have separated the general and SC classes from the test set by green and red points, respectively, in Figure 3(b) by a non-linear projection from a high-dimensional feature space to 2-D and 3-D images (displays) using t-SNE [24]. The two classes are highly overlapped. Therefore, COVID-19 severity classification can be regarded as a non-linear fine-grained pattern recognition problem [25]. We thus sought assistance from NN learning [26] for strong non-linear categorisation capacity, and thereby established a novel NN model with three hidden layers with a delicately designed softmax loss functions, as shown in Figure 3(a), within which the inter-class distances are substantially stretched. Inspiringly, with the features sequentially extracted by each layer, the boundaries of these two severity classes gradually surface, as shown in Figure 3(b)–(d) with stretched discrimination margins. It can be observed they are naturally not linearly separable, but rather highly overlapping. Therefore, the present neural network model is a niche ML manner that benefits from clear interpretability potential due to its visualised learning evolution attained sequentially through the hidden layers in Figure 3(a).

The two-way classification (general VS SC) performance comparison among the proposed A-softmax NN model and the other classifiers is listed in Table 2, using the same ‘30 + 1’ features. On all the evaluation metrics, A-softmax NN model yields the best overall performance consistently, which demonstrates the superiority of our proposition. Meanwhile, the acquired overall accuracy (i.e. 86.30%) and F1-socre (i.e. 0.8195) is also high. This verifies the differentiation capability of the ‘30 + 1’-feature combined indicator by the present neural learning model for severity classification of COVID-19 patients. It is impressive that, for SC (severe + critical) class our method can achieve a high recall rate of 0.9050 with the precision of 0.9113. Thus, it is revealed that, the present approach can screen SC patients out effectively to facilitate early triage, diagnosis, and treatment, which is essential for the practical applications in spirit of saving patients of high death risk timely. On general class, the performance of our approach is relatively low with F1-score of 0.7309, but still better than the others. This may be due to the three main issues: (1) the high pattern overlap as in Figure 3(b) and (d), (2) due to the retrospective study limitation, the used 2139 samples actually suffer from the feature missing problem although ‘0’ padding is used, and (3) the training sample size (about 1711 per test fold) does not suffice exploration of the non-linear pattern discriminative capacity of neural network. The detailed confusion matrix of A-softmax NN model is given in Table 3. Linear SVM is the state-of-the-art linear classifier, which has been demonstrated inferior to non-linear A-softmax NN in all the test cases. Therefore, COVID-19 severity classification is a non-linear fine-grained pattern recognition problem as in Figure 3. Thus, designing and training an efficient non-linear classifier become an indispensable way to fulfill this challenging classification mission.

More importantly, the present A-softmax NN provides a balance between model interpretability and categorisation accuracy. Bearing in mind that clinical settings tend to prefer interpretable models, this research could be expected to pave the way from a black box NN model to interpretable severity classification application.

**Ablation study 1: selected laboratory indicator number**

To investigate the effect of selected laboratory indicator number towards severity classification of COVID-19 patients, we set it to 5, 10, 20, 30, 40, 50, 60, and 70, respectively, plus the age feature. The performance comparison among the different feature dimensionalities is exhibited in Table 4. Particularly, selecting 30 laboratory indicators as we do can achieve the best performance in five test cases, especially on overall accuracy, recall on SC class and F1-score on SC class.

Significantly, only using five laboratory indicators can also achieve promising results. While, we need to clarify that it is still hard to judge what the optimal feature dimensionality is. The main reason is that, due to the retrospective study limitation, the used 2139 samples actually suffer from the serious feature missing problem. For example, the feature missing rate for dimensionality 30 is about 25%. Meanwhile, the used ‘0’ padding strategy may weaken performance unpredictably. As a consequence, one cannot judge the optimal laboratory indicator number of a high confidence according to the current experimental results. Interestingly, just selecting 30 laboratory indicators ensures satisfactory classification performance with sufficiently low overfitting.

**Ablation study 2: training batch size**

To investigate the effect of batch size for NN training towards severity classification of COVID-19 patients, we set it to 16,
32, 64, 128, and 256, respectively. The performance comparison among them is listed in Table 5. Generally, the results are not greatly sensitive to training batch size. The batch sizes of 16, 32, 64, and 128 are of similar performance.

### Ablation study 3: learning rate

To investigate the effect of learning rate for NN training towards severity classification of COVID-19 patients, we set it to 0.0001, 0.001, 0.01, and 0.1, respectively. Their performance comparison is listed in Table 6. Overall, the learning rate of 0.001 chosen by us can achieve the best performance in most of the test cases.

### Ablation study 4: weight decay

To investigate the effect of weight decay for NN training towards severity classification of COVID-19 patients, we set it to 0.0001, 0.001, 0.005, and 0.01, respectively. The performance comparison among them is listed in Table 7. Overall, the weight decay is picked to achieve the best performance in most of the cases. While the results are not greatly sensitive to weight decay. The weight decays of 0.0001, 0.001, 0.005, and 0.01 are of similar performance.

### Ablation study 5: neural network optimiser

To investigate the effect of weight decay for NN training towards severity classification of COVID-19 patients, we compare the performance of two widely used optimisers:
### Table 4: Fine-grained two-class identification (general VS SC) performance comparison among the different feature dimensionalities. For each case, the best performance is shown in boldface.

| Feature dimensionality | Test case | Precision | Recall | F1-score | Accuracy |
|------------------------|-----------|-----------|--------|----------|----------|
| 5 + 1                  | General   | 0.7150    | 0.8130 | 0.7608   |          |
|                        | SC        | 0.9338    | 0.8906 | 0.9117   |          |
|                        | Overall   | 0.8244    | 0.8518 | 0.8363   | 87.10%   |
| 10 + 1                 | General   | 0.7339    | 0.8033 | 0.7671   |          |
|                        | SC        | 0.9316    | 0.9019 | 0.9165   |          |
|                        | Overall   | 0.8327    | 0.8526 | 0.8418   | 87.70%   |
| 20 + 1                 | General   | 0.7263    | 0.7808 | 0.7525   |          |
|                        | SC        | 0.9262    | 0.9033 | 0.9146   |          |
|                        | Overall   | 0.8262    | 0.8421 | 0.8336   | 87.30%   |
| 30 + 1                 | General   | 0.7236    | 0.7384 | 0.7309   |          |
|                        | SC        | 0.9113    | 0.9050 | 0.9081   |          |
|                        | Overall   | 0.8175    | 0.8217 | 0.8195   | 86.30%   |
| 40 + 1                 | General   | 0.6937    | 0.7361 | 0.7143   |          |
|                        | SC        | 0.9114    | 0.8931 | 0.9022   |          |
|                        | Overall   | 0.8026    | 0.8146 | 0.8082   | 85.42%   |
| 50 + 1                 | General   | 0.7094    | 0.7231 | 0.7162   |          |
|                        | SC        | 0.9087    | 0.9030 | 0.9058   |          |
|                        | Overall   | 0.8091    | 0.8130 | 0.8110   | 85.86%   |
| 60 + 1                 | General   | 0.7059    | 0.7006 | 0.7032   |          |
|                        | SC        | 0.9006    | 0.9028 | 0.9017   |          |
|                        | Overall   | 0.8032    | 0.8017 | 0.8025   | 85.23%   |
| 70 + 1                 | General   | 0.7035    | 0.7208 | 0.7120   |          |
|                        | SC        | 0.9064    | 0.8991 | 0.9027   |          |
|                        | Overall   | 0.8050    | 0.8099 | 0.8074   | 85.46%   |

Abbreviation: SC, severe + critical.

### Table 5: Fine-grained two-class identification (general VS SC) performance comparison among the different training batch sizes. For each case, the best performance is shown in boldface.

| Training batch size | Test case | Precision | Recall | F1-score | Accuracy |
|---------------------|-----------|-----------|--------|----------|----------|
| 16                  | General   | 0.7407    | 0.7050 | 0.7224   |          |
|                     | SC        | 0.9022    | 0.9169 | 0.9095   |          |
|                     | Overall   | 0.8215    | 0.8109 | 0.8160   | 86.35%   |
| 32                  | General   | 0.7236    | 0.7384 | 0.7309   |          |
|                     | SC        | 0.9113    | 0.9050 | 0.9081   |          |
|                     | Overall   | 0.8175    | 0.8217 | 0.8195   | 86.30%   |
| 64                  | General   | 0.7193    | 0.7607 | 0.7394   |          |
|                     | SC        | 0.9178    | 0.9000 | 0.9088   |          |
|                     | Overall   | 0.8185    | 0.8303 | 0.8241   | 86.49%   |
| 128                 | General   | 0.7040    | 0.7811 | 0.7405   |          |
|                     | SC        | 0.9234    | 0.8894 | 0.9061   |          |
|                     | Overall   | 0.8137    | 0.8352 | 0.8233   | 86.21%   |
| 256                 | General   | 0.6898    | 0.8126 | 0.7462   |          |
|                     | SC        | 0.9328    | 0.8769 | 0.9040   |          |
|                     | Overall   | 0.8113    | 0.8447 | 0.8251   | 86.07%   |

Abbreviation: SC, severe + critical.
 Abbreviation: SC, severe + critical.

Adam and SGD [27]. The performance comparison between them is listed in Table 8. Overall, Adam outperforms SGD in most of the test cases. While the performance gap is not significant.

### TABLE 7  Fine-grained two-class identification (general VS SC) performance comparison among the different weight decays. For each case, the best performance is shown in boldface

| Weight decay | Test case | Precision | Recall | F1-score | Accuracy |
|--------------|-----------|-----------|--------|----------|----------|
| 0.0001       | General   | 0.7246    | 0.7273 | 0.7259   |          |
| SC           | 0.9080    | 0.9069    | 0.9074 |          |          |
| Overall      | 0.8163    | 0.8171    | 0.8167 | 86.16%   |          |
| 0.0005       | General   | 0.7123    | 0.7440 | 0.7278   |          |
| SC           | 0.9124    | 0.8988    | 0.9055 |          |          |
| Overall      | 0.8123    | 0.8214    | 0.8167 | 85.97%   |          |
| 0.001        | General   | 0.7236    | 0.7384 | 0.7309   |          |
| SC           | 0.9113    | 0.9050    | 0.9081 |          |          |
| Overall      | 0.8175    | 0.8217    | 0.8195 | 86.30%   |          |
| 0.005        | General   | 0.7209    | 0.7236 | 0.7222   |          |
| SC           | 0.9068    | 0.9056    | 0.9062 |          |          |
| Overall      | 0.8138    | 0.8146    | 0.8142 | 85.97%   |          |
| 0.01         | General   | 0.7218    | 0.7365 | 0.7291   |          |
| SC           | 0.9106    | 0.9044    | 0.9075 |          |          |
| Overall      | 0.8162    | 0.8205    | 0.8183 | 86.21%   |          |

Abbreviation: SC, severe + critical.

### TABLE 8  Fine-grained two-class identification (general VS SC) performance comparison between different neural network optimisers. For each case, the best performance is shown in boldface

| Optimiser  | Test case | Precision | Recall | F1-score | Accuracy |
|------------|-----------|-----------|--------|----------|----------|
| Adam       | General   | 0.7236    | 0.7384 | 0.7309   |          |
| SC         | 0.9113    | 0.9050    | 0.9081 |          |          |
| Overall    | 0.8175    | 0.8217    | 0.8195 | 86.30%   |          |
| SGD        | General   | 0.7310    | 0.7310 | 0.7310   |          |
| SC         | 0.9094    | 0.9094    | 0.9094 |          |          |
| Overall    | 0.8202    | 0.8202    | 0.8202 | 86.44%   |          |

Abbreviation: SC, severe + critical.

### Ablation study 6: A-softmax NN training convergence

To investigate whether the training procedure of A-softmax NN can converge with 60 training epochs, we list the average performance of A-softmax NN model over 5-fold training sets with increasing training epoch in Figure 4. Overall, the present NN converges when training iteration reaches 60 epochs. Particularly, all the four evaluation metrics go beyond 0.9. This verifies the fitting capacity of A-softmax NN model on training set for COVID-19 severity characterisation.

### DISCUSSION

Machine intelligence has the potential to assist physicians and healthcare professionals to derive early diagnosis and therapy, through rapid severity classification using data-rich clinical and laboratory features. COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has distinct epidemiological and clinical characteristics.

![Figure 4](image-url)
Accordingly, a number of critical laboratory features, used as indicators, can be used to help identify and predict the prognosis of patients. To date, the majority of retrospective studies performed have been done using small patient cohorts with significant limitations to predict disease outcome. The clinical characteristics of COVID-19 remain unclear, in particular the clinical course of the disease. In this study, with the assistance of ML techniques, we sought to classify 2662 COVID-19 cases into two categories that is general and severe or critical. To accomplish this classification, a five-layer NN model was established to find the top 30 features as a more precise classification indicator with clear clinical interpretability, to provide timely and effective personalised treatment.

By non-linear projection going from high-to-low-dimensional space features, the two designated COVID-19 severity levels can be extremely mixed, thus complicating disease differentiation. Physicians typically classify patients using clinical guidelines for diagnosis and treatment following admission [8]. However, beyond respiratory symptoms, other factors are present that influence patient status, including hypertension, diabetes, coronary heart disease, malignant tumours, chronic pulmonary disease, immunosuppression, and so on. The co-existence of these other conditions can contribute to the degree of illness, causing the deterioration of moderate cases into a severe or critical classification. Therefore, in the early stages of COVID-19, early diagnosis with fine classification and individualised treatment are essential to the prognosis of the patient. Laboratory clinical features may be more objective and sensitive indexes to be implemented. Based on the well-accepted clustering standard of inter-class deviation over intra-class deviation, this study demonstrates a multi-feature combined indicator can minimise intra-class confusion and more clearly reflect the functional status of multiple organs [15].

Herein, Hs-CRP and LDH rank the top-2 highest among the top-30 features, which is consistent with most of the recent reports, that Hs-CRP and LDH are two of the most important factors for early warning of severe or critically ill patients with COVID-19. In a large retrospective cohort study among patients with COVID-19 in China, leucocytosis, d-d dimer quantification, myoglobin, interleukin 2 receptor, and ferritin were also correlated with the severe/critical classification and early warning of potential mortality [28]. These observations are highly consistent with the top-30 features identified in this study. Moreover, these 30 indicators also include the assessment of inflammatory response, tissue perfusion, thyroid function, blood coagulation, liver and kidney function, electrolyte, and acid-base balance, which can be accurately used in the severity classification of COVID-19 patients. Along with the top-30 clinical features identified, age is also an indispensable supplementary feature with a strong relationship to disease severity, which has also been confirmed in other reports [1–3].

Due to the rapid progression of COVID-19 illness, potentially severe and critical cases must be identified early and treated in the appropriate settings as soon as possible. To accomplish this objective, we have designed a sensitive Softmax loss function in the present NN model, in which the inter-class distance has been substantially stretched, sharpening the class boundaries. This neural network model yields classification results with zero inconsistent rate and a ms-level of computational cost. For two-way classification (i.e. general VS SC), the blind test F1-score for SC and general are 0.9081 and 0.7309. The latter is due to the serious overlaps of SC and general cases (Figure 3(b)–(d)), complicating the two-way classification especially at the early stage. This observation is consistent with what we have observed in clinical practice. Patients with oxygen saturation below 93% are defined as severe patients. Among them, patients with type I respiratory failure as critical cases needed to be treated with ventilator support. After admission, many critically ill patients died due to the delayed ventilator support. To date, the need and appropriate timing for non-invasive mechanical ventilation (NMV) or invasive mechanical ventilation (IMV) for critical patients has received increasing attention among medical workers [29]. On the other hand, when residents treated patients with complications, they tended to consider patients’ features more comprehensively yet promptly before providing appropriate treatments. In order to determine the error induced by this common objective, we established the NN model for two-way classification (i.e. general VS SC). In brief, this model achieved 86.30% blind test accuracy, 0.8195 F1-score, and 100% consistency. With the high recall rate of 0.9050 and precision of 0.9113 in the SC population, the effectiveness and the interpretability of the present combination index has been verified. Therefore, our classification model could assist physicians with a high degree of precision to identify the severe COVID-19 patients with 100% consistency to significantly improve diagnosis accuracy and provide early therapeutic intervention to improve the prognosis of patients and relieve the growing pressures due to late-stage diagnoses of COVID-19 on healthcare services.

Our study does have a number of limitations. First, based on the retrospective study design, not all laboratory tests were performed on all patients. As a consequence, some other key factors may not have been included in the model, thus potentially underestimating case classification for in-hospital patients. Second, the qualitative clinical indicator and personal treatment information has not been involved in AI decision due to the quantisation difficulty. The classification could be improved by the incorporation of these features with the development of data processing technology in future. Undoubtedly, the classification accuracy could be further enhanced if more physicians are involved in supervisory training to reduce subjective bias. In addition, categorisation performance and interpretability of our findings might be limited by sample size. Finally, only single centre retrospective study is conducted now, which provides a preliminary assessment of the clinical course and outcome of patients. More abundant samples and multi-centred study in future will be expected to improve the generality of the present results.
In conclusion, our AI model provides a framework for clinicians and public health workers to accurately evaluate the severity of COVID-19 patients in the early stage of their illness so as to advance the time window of individualised treatment for achieving better clinical outcomes.

ACKNOWLEDGEMENTS
The authors dedicate the study to those who have devoted their lives to the battle with COVID-19.

This work was jointly supported by the COVID-19 Prompt Response Research Special Project from Huazhong University of Science and Technology under grants. 2020kyfXGY]113 and 2020kyfXGY]023, a special fund for novel coronavirus pneumonia from the Science and Technology Department, Hubei province, under grant 2020FC0A35 and Wuhan Science and Technology Bureau Foundation under grant 20170602010 10161. The funding sources had no role in the study design, data collection, analysis, interpretation, or writing of the report.

ORCID
Hai-Tao Zhang https://orcid.org/0000-0002-8819-8829

REFERENCES
1. Wang, D., et al.: Clinical characteristics of 138 hospitalised patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. J. Am. Med. Assoc. 323(11), 1061–1069 (2020). https://doi.org/10.1001/jama.2020.1585
2. Chen, N., et al.: Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 395(10223), 507–513 (2020)
3. Guan, W., et al.: Clinical characteristics of coronavirus disease 2019 in China. N. Engl. J. Med. 382, 1708–1720 (2020). https://doi.org/10.1056/NEJMoA2002052
4. Wu, J.T., Leung, K., Bushman, M.: Estimating clinical severity of covid-19 from the transmission dynamics in Wuhan, China. Nat. Methods. 26, 506–510 (2020). https://doi.org/10.1038/s41592-020-0822-7
5. Yu, H., et al.: Data-driven discovery of clinical routes for severity detection in covid-19 paediatric cases. medRxiv (2020). https://doi.org/10.1101/2020.03.09.20032219
6. Yan, L., et al.: A machine learning-based model for survival prediction in patients with severe covid-19 infection. medRxiv (2020). https://doi.org/10.1101/2020.02.27.20028027
7. Munster, V.J., et al.: A novel coronavirus emerging in China - key questions for impact assessment. N. Engl. J. Med. 382, 692–694 (2020). https://doi.org/10.1056/NEJMmp2000929
8. National Health Commission of the People’s Republic of China: Diagnosis and treatment of pneumonia infected by the new novel coronavirus (trial 5th edition). The Medical Letter from the National Health Office, Beijing, China (2020)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT Press, Cambridge, MA (2016)
10. Pereira, R.M., et al.: COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Comput. Methods Progr. Biomed. 194, 105532 (2020)
11. Sun, L., et al.: Adaptive feature selection guided deep forest for covid-19 classification with chest CT. IEEE J. Biomed. Health Inf. https://doi.org/10.1109/JBHI.2020.3019505
12. Harmon, S.A., et al.: Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat. Commun. 11(1), 1–7 (2020)
13. Yao, H.C., et al: Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. Front. Cell Dev. Biol. 8, 684 (2020)
14. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 1–27 (2011)
15. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification, 2nd ed. Wiley, New York, NY (2001)
16. Xu, B., et al.: Empirical evaluation of rectified activations in convolutional network. arXiv:1505.08853 (2015)
17. Liu, W., et al.: Deep hypersphere embedding for face recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 1 (2017)
18. Diederik, K.P., Adam, J.B.: A method for stochastic optimization. arXiv:1412.6980 (2014)
19. Powers, D.M.: Evaluation: from precision, recall and F-score to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2(1), 37–63 (2011)
20. Yuan, Y., et al.: Data driven discovery of cyber physical systems. Nat. Commun. 10(1), 1–9 (2019)
21. Rennie, J., et al.: Tackling the poor assumptions of naive Bayes classifiers. In: Proc. International Conference on machine learning (ICML), pp. 1449–1457 (2003)
22. Altman, N.S.: An introduction to kernel and nearest-neighbour nonparametric regression. Am. Stat. 46(3), 175–185 (1992)
23. Srivastava, S., Gupta, M.R., Frigylik, B.A.: Bayesian quadratic discriminant analysis. J. Mach. Learn. Res. 8(Jun), 1277–1305 (2007)
24. van der Maaten, L.J.P., Hinton, G.E.: Visualising high-dimensional data using t-sne. J. Mach. Learn. Res. 9, 2579–2605 (2008)
25. Lin, T.Y., Roychowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In:Proc. IEEE International Conference on computer vision (ICCV), pp. 1449–1457 (2015)
26. Kohonen, T.: An introduction to neural computing. Neural Net. 1(1), 3–16 (1988)
27. Li, Y., et al.: Pipe-sgd: a decentralised pipelined sgd framework for distributed deep net training. In:Proc. Advances in neural information processing systems (NIPS), pp. 8045–8056 (2018)
28. Zhou, F., et al.: Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 395(10229), 1054–1062 (2020). https://doi.org/10.1016/S0140-6736(20)30566-3
29. Yan, H.-Y., Yang, Y., Wu, Y.-L.: Clinical analysis of optimal timing for application of noninvasive positive pressure ventilation in treatment of AECOPD patients. Eur. Rev. Med. Pharmacol. Sci. 18(15), 2176–2181 (2014)

How to cite this article: Xiao Y, Yan L, Zhang M, et al. Machine learning discovery of distinguishing laboratory features for severity classification of COVID-19 patients. IET Cyber-Syst. Robot. 2021;3:31–43. https://doi.org/10.1049/ey2.12005