ORIGINAL RESEARCH

Machine Learning-Based Decision Model to Distinguish Between COVID-19 and Influenza: A Retrospective, Two-Centered, Diagnostic Study

Xianlong Zhou[1,]*
Zhichao Wang [2,]*
Shaoping Li[1]
Tanghai Liu[3]
Xiaolin Wang [4]
Jian Xia[1]
Yan Zhao[1]

[1]Emergency Center, Hubei Clinical Research Center for Emergency and Resuscitation, Zhongnan Hospital of Wuhan University, Wuhan, Hubei, 430071, People's Republic of China; [2]Emergency Department, Wuhan No. 1 Hospital, Wuhan, Hubei, 430022, People's Republic of China; [3]Information Center, Wuhan No. 1 Hospital, Wuhan, Hubei, 430022, People's Republic of China; [4]Gennlife (Beijing) Biotechnology Co. Ltd, Haidian, Beijing, 100080, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yan Zhao
Emergency Center, Hubei Clinical Research Center for Emergency and Resuscitation, Zhongnan Hospital of Wuhan University, Wuhan, Hubei, 430071, People's Republic of China
Email doctoryanzhao@whu.edu.cn

Jian Xia
Emergency Center, Hubei Clinical Research Center for Emergency and Resuscitation, Zhongnan Hospital of Wuhan University, Wuhan, Hubei, 430071, People's Republic of China
Email jianjian_1998@sina.com

Background: Considering the current situation of the novel coronavirus disease (COVID-19) epidemic control, it is highly likely that COVID-19 and influenza may coincide during the approaching winter season. However, there is no available tool that can rapidly and precisely distinguish between these two diseases in the absence of laboratory evidence of specific pathogens.

Methods: Laboratory-confirmed COVID-19 and influenza patients between December 1, 2019 and February 29, 2020, from Zhongnan Hospital of Wuhan University (ZHWU) and Wuhan No.1 Hospital (WNH) located in Wuhan, China, were included for analysis. A machine learning-based decision model was developed using the XGBoost algorithms.

Results: Data of 357 COVID-19 and 1893 influenza patients from ZHWU were split into a training and a testing set in the ratio 7:3, while the dataset from WNH (308 COVID-19 and 312 influenza patients) was preserved for an external test. Model-based decision tree selected age, serum high-sensitivity C-reactive protein and circulating monocytes as meaningful indicators for classifying COVID-19 and influenza cases. In the training, testing and external sets, the model achieved good performance in identifying COVID-19 from influenza cases with a corresponding area under the receiver operating characteristic curve (AUC) of 0.94 (95% CI 0.93, 0.96), 0.93 (95% CI 0.90, 0.96), and 0.84 (95% CI: 0.81, 0.87), respectively.

Conclusion: Machine learning provides a tool that can rapidly and accurately distinguish between COVID-19 and influenza cases. This finding would be particularly useful in regions with massive co-occurrences of COVID-19 and influenza cases while limited resources for laboratory testing of specific pathogens.

Keywords: COVID-19, influenza, classification, machine learning, diagnostic accuracy

Background

The outbreak of the coronavirus disease 2019 (COVID-19), which is caused by a novel coronavirus known as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has emerged as a severe global health problem.[1] As of October 1, 2020, the COVID-19 outbreak was reported to have affected 33,842,281 individuals, including 1,010,634 deaths worldwide.[2] Numerous experts have warned that a second wave of COVID-19 could considerably be more devastating because it is likely to coincide with the start of the 2020–2021 winter influenza season.[3]

Because COVID-19 and influenza display significant similarities in their transmission routes and symptoms, it is a challenge to distinguish between these two respiratory diseases, particularly in the early stage, based on a common diagnosis in

the absence of laboratory evidence of specific pathogens.[4] Although our previous study indicated that COVID-19 likely interfered with influenza,[5] these two diseases could coincide in space and time in the approaching winter months based on the current situation of the epidemic control. To date, no reliable tool based on simple variables is yet available for providing a differential diagnosis between COVID-19 and influenza. In the present study, the retrospective data of COVID-19 and influenza patients from two large teaching hospitals in Wuhan, China were used with a machine learning based-modeling approach for developing a tool devoted to accurately classify patients in the corresponding disease.

Methods
Study Design and Participants
This retrospective study was performed in Zhongnan Hospital of Wuhan University (ZHWU) and Wuhan No.1 Hospital (WNH) located in Wuhan, China. All laboratory-confirmed COVID-19 and influenza patients from ZHWU and WNH between December 1, 2019, and February 29, 2020, were eligible for analysis. We chose to focus on the cases infected only during this period because during the past two influenza seasons, the influenza incidence markedly increased in December and achieved peaks in January of the following year in Wuhan City.[5] Considering the possible co-circulation of COVID-19 and influenza in Wuhan, we decided to focus on studying only those infected cases between December 1, 2019, and February 29, 2020. Patients who had incomplete medical records, any other infections including the co-infection of SARS-CoV-2 and influenza virus (type A and/or B), immunosuppression, malignancies, pregnancy, and those who received any treatment prior to their visit to the emergency department or outpatient clinic were excluded. The demographic, clinical, and laboratory data of the COVID-19 and influenza patients were retrieved from electronic medical records.

Case Definition
A laboratory-confirmed case of COVID-19 was defined as a suspected case with laboratory evidence of the SARS-CoV-2 infection detected by real-time reverse-transcription–polymerase-chain-reaction (RT-PCR). Throat-swab RNA was extracted and tested by real-time RT-PCR with SARS-CoV-2 specific primers and probes. Two target genes, including open reading frame 1ab (ORF1ab) and

nucleocapsid protein (N), were simultaneously amplified and tested during the real-time RT-PCR. The real-time RT-PCR assay was performed using a SARS-CoV-2 nucleic acid detection kit (DAAN, Guangzhou, China) according to the recommendation of the World Health Organization (WHO).[6] A laboratory-confirmed case of influenza was defined as the influenza-like illness case with the laboratory evidence of influenza virus infection (type A and/or B) by a rapid detection of influenza viral antigens. The influenza A and B viral antigens in the throat swabs were determined by a commercial flu A&B test kit (Wondfo, Guangzhou, China) in accordance with the National Protocol of Influenza Surveillance.[7]

Development of a Machine Learning-Based Model
The COVID-19 and influenza data from ZHWU were divided into a training set and a testing set in the ratio 7:3. The demographic data, life style, comorbidities, physical signs and symptoms at the time of the hospital visit, and all available data of laboratory testing data obtained in the outpatient clinic or emergency department were considered as potential features for the model development. Features were excluded for subsequent analysis whenever the missing data accounted for more than 20%.

The prediction model was based on the XGboost machine learning algorithm, which has been favorably assessed according to a substantial number of features when compared to concurrent approaches, including a remarkable interpretability potential due to its recursive tree-based decision system.[8] Training was made using the following XGBoost parameter values: maximum depth (max_depth) = 4, learning rate (eta) = 0.2, regularization parameter alpha (α) = 1, subsample = 0.9, colsample_by-tree = 0.9, objective = 'binary: logistic', number of rounds (nrounds) = 50, and number of tree estimators (n_estimators) = 150. This algorithm was named as the "multi-tree XGBoost".

Determination of Key Features
To determine the key features for the decision model, the contribution of each feature to the algorithm's decision was evaluated. The top 10 important features were ranked based on their relative importance in the multi-tree XGBoost algorithm as previously described.[9] Briefly, we first selected 100 random number seeds from 0 to 99, and the data for each number seed were divided into training

and testing sets in the ratio 7:3. Then, a multi-tree XGBoost was trained for each number seed, and the average importance of each feature could be determined with the ensured stability of the feature rankings.

Subsequently, a 5-fold cross-validation method was used to determine the key features for the model development. The selection of key features was based on a procedure assessing the area under the receiver operating characteristic curve (AUC) score of an increasing number of features as follows: the top feature was first used for prediction, and the average AUC scores of the training and the testing sets were calculated. Then, the top two to top ten features were added to the top feature in sequence, and the corresponding AUC scores were examined according to the same sequence: features were considered as valuable additional key features for the decision tree until the corresponding relative increase of AUC score was below the threshold of 1%.

Development of a Feasible Decision Tree

To establish a clinically interpretable decision tree, the number of tree has been reduced to 1, which leads to the "single-tree XGBoost". The single-tree XGBoost was trained using selected key features as previously described[9] The parameters for the single-tree XGBoost model training can be summarized as follows: max_depth = 4, eta = 0.2, α = 0, subsample = 1, colsample_bytree = 1, objective='binary: logistic', nrounds = 50, and n_estimators = 1. The specificity, sensitivity, NPV, PPV, accuracy and AUC scores were calculated to evaluate the performance of prediction. Finally, the structure of a clinically interpretable decision tree with reduced complexity was obtained by the split of all COVID-19 and influenza patients after data imputation.

External Test

In order to assess the performance of the decision model with an external test, the single-tree XGBoost algorithm was applied to the external dataset from WNH, and the corresponding specificity, sensitivity, accuracy, NPV, PPV, and AUC scores were calculated.

Statistical Analysis

The characteristics of the COVID-19 and influenza patients were compared using the Mann–Whitney U-test or the Chi-square test wherever appropriate. Two-sided *P* values of less than 0.05 were considered to indicate statistical significance. Model output was the prediction of the type of disease assigned to each case, ie, either COVID-19 or influenza, with label 1 and label 0 arbitrarily assigned to COVID-19 and influenza status (ie, positive cases and negative cases), respectively. The diagnostic performance of the model was evaluated using sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV) and accuracy. Accuracy was defined as the ratio of (TP + TN)/(TP+FP+TN+FN), and TP, TN, FP, FN stand for true positive, true negative, false positive and false negative, respectively. All parameters were calculated and presented with two-sided 95% confidence interval (CI). In addition, the performance of the model was compared with logistic regression, which has been considered as a standard method. Multiple imputation was used to handle the problem of the missing values for features in both model development and external validation. The performance of the proposed model was also tested using the datasets containing influenza type A or type B cases. The generalizability of the prediction model was evaluated on the external dataset. All the analyses were performed with the use of the R software (version 4.0.2, R Foundation for Statistical Computing).

Results

This study was performed between December 1, 2019, and February 29, 2020, in ZHWU and WNH. We obtained the data of 357 COVID-19 and 1893 influenza (including 1291 type A and 602 type B cases) patients from ZHWU for model training and testing, and the external validity of the model was assessed on the external validation dataset from WNH, comprised of 308 COVID-19 and 312 influenza patients (Figure 1). In the model development dataset, the median age of the COVID-19 patients was 58 years (IQR 44–70), whereas the influenza patients had a median age of 8 years (IQR 5–12) (*p* < 0.01). No significant difference was observed in the sex distributions of COVID-19 and influenza patients (male, 49.6% vs 54.8%, *p* = 0.08). Fever was the most common symptom for both the COVID-19 (85.5%) and influenza (97.1%) patients (*p* < 0.01). In the external test dataset, the median age of COVID-19 was also significantly higher than that of influenza cases (62 [IQR 47–70] vs 13 [IQR 8–31] years, *p* < 0.01). The results of sex distributions and symptoms were similar as that of the model development dataset. Clinical signs and symptoms, laboratory data including blood cell counts, blood biochemistry, coagulation function, and the infection markers of the
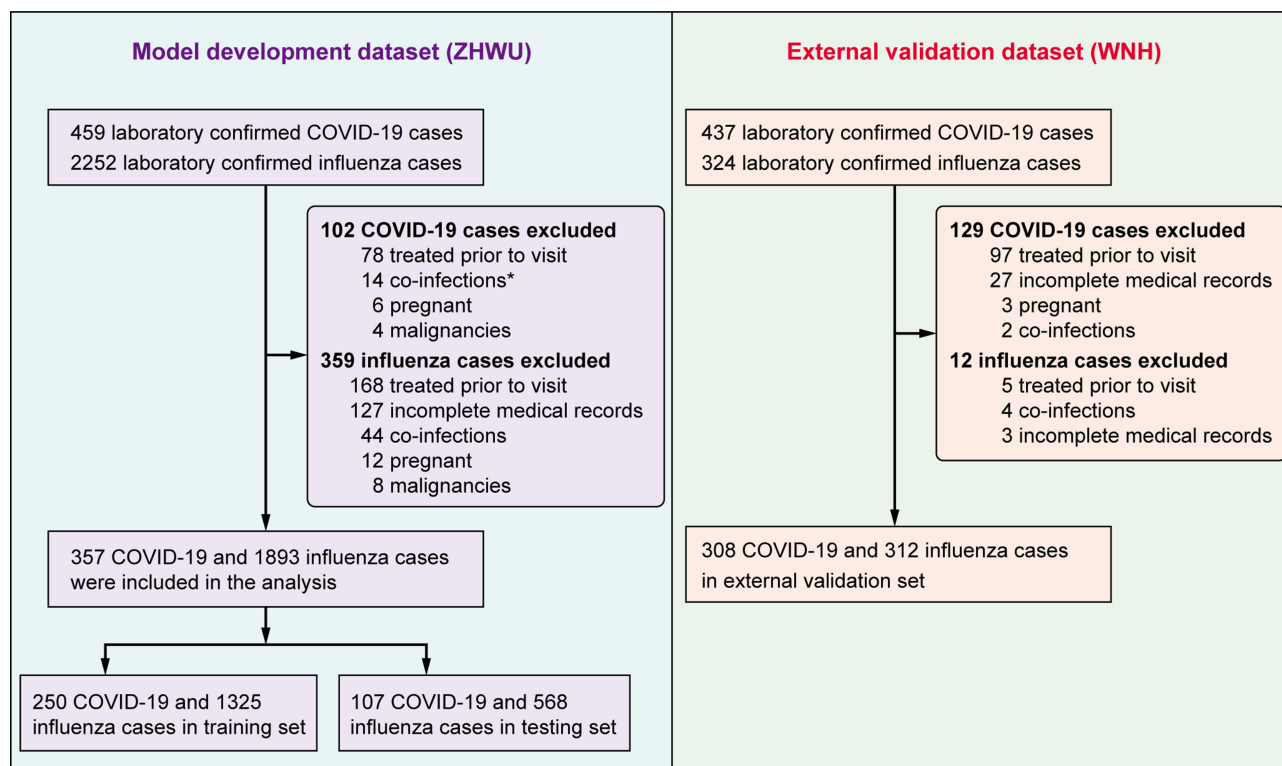
Figure 1 Study profile. Data from ZHWU were used for model development and data from WNH were used for external validation. *Co-infections indicates any other infection and those who had both SARS-CoV-2 and influenza virus infections.

patients are summarized in Table 1 and supplementary materials (Table S1).

In total, 110 features were initially included as potential indicators after considering the corresponding feasibility and timeliness. There were 52 features excluded because of too many missing data, and the remaining 58 features were included for subsequent analysis (Table S2). The importance of the features considered in the prediction model was ranked by multi-tree XGBoost (Figure S1). Age feature had the greatest impact on prediction, and serum hsCRP was the second most important feature contributing to the decision model. The cell count of the monocytes was ranked as the third most important feature and the mean corpuscular hemoglobin concentration (MCHC) was ranked as the fourth. No vital signs or symptoms were ranked as important indicators for prediction. Based on the relative importance for the decision model, the age feature was used first for COVID-19 prediction. The AUC scores for the training and testing sets were 0.92 (95% CI 0.92, 0.92) and 0.91 (95% CI 0.91, 0.92), which indicated that age is crucial for the classification of COVID-19 and influenza patients. The performance of the model indicated no marked improvement in

AUC scores (+0.93% in training set) when the number of features was increased from three (age, hsCRP and monocytes) to four (age, hsCRP, monocytes and MCHC) (Table 2, Table S3, Figure S2). Finally, the proposed machine-learning model was developed using age, hsCRP, and monocytes.

The performance of the single-tree XGBoost model is presented in Figure 2 and Table 3. The proposed model was developed using the data of 357 COVID-19 and 1893 influenza patients from ZHWU. For the training set, the sensitivity was 0.91 (95% CI 0.87, 0.94), specificity was 0.98 (95% CI 0.97, 0.99) and the accuracy for the COVID-19 prediction was 0.97 (95% CI 0.97, 0.97). The testing set achieved a sensitivity of 0.88 (95% CI 0.820, 0.942), a specificity of 0.98 (95% CI 0.96, 0.99), and an accuracy of 0.96 (95% CI 0.96, 0.96). In addition, the external test demonstrated a prediction accuracy of 0.84 (95% CI 0.84, 0.84), with a sensitivity of 0.91 (95% CI 0.87, 0.94) and a specificity of 0.77 (95% CI 0.73, 0.82). The AUC scores of the training, testing and external test sets were 0.94 (95% CI 0.93, 0.96), 0.93 (95% CI 0.90, 0.96) and 0.84 (95% CI 0.81, 0.87), respectively. The confusion matrix are summarized in the supplementary materials (Figure S3, S4, S5). Moreover, the

Table 1 Characteristics of Patients Recruited for Model Development and External Validation

Characteristics	ZHWU (n=2250)		p value	WNH (n=620)		p value
	COVID-19 (n=357)	Influenza (n=1893)		COVID-19 (n=308)	Influenza (n=312)	
Demographic						
Age, years	58 [44, 70]	8 [5, 12]	<0.01	62 [47, 70]	13 [8, 31]	<0.01
Gender, male, n (%)	177 (49.6)	1038 (54.8)	0.08	157 (51.0)	170 (54.5)	0.42
Clinical data						
Vital signs						
SBP (mmHg, ref: 90–140)	124 [116, 138]	118 [102, 123]	0.06	120 [112, 135]	121 [104, 128]	0.08
DBP (mmHg, ref: 60–90)	75 [69, 82]	74 [72, 81]	0.09	73 [65, 80]	74 [69, 82]	0.12
RR (bpm, ref: 12–20)	21 [20, 23]	21 [20, 24]	0.25	23 [18, 25]	21 [21, 24]	0.33
SpO_2 (%, ref: ≥ 94)	96 [94, 98]	96 [95, 98]	0.83	95 [93, 98]	94 [93, 98]	0.66
Temperature (°C, ref: 36.2–37.3)	37.8 [37.5, 38.5]	38.2 [37.0, 38.8]	0.39	38.1 [37.5, 38.2]	38.3 [37.0, 38.5]	0.17
Symptoms						
Fever, n (%)	305 (85.4)	1838 (97.1)	<0.01	255 (82.8)	267 (85.6)	0.38
Fatigue, n (%)	197 (55.2)	1021 (53.9)	0.69	162 (52.6)	168 (53.8)	0.81
Cough, n (%)	176 (49.3)	1300 (68.7)	<0.01	154 (50.0)	183 (58.7)	0.04
Myalgia, n (%)	89 (24.9)	462 (24.4)	0.84	62 (20.1)	80 (25.6)	0.11
Headache, n (%)	70 (19.6)	982 (51.9)	<0.01	68 (22.1)	69 (22.1)	1.00
Laboratory findings						
Blood cell counts						
Leucocytes ($1×10^9$/L, ref: 3.5–9.5)	5.48 [4.18, 7.20]	7.10 [5.74, 8.76]	<0.01	5.82 [4.37, 6.78]	6.57 [5.12, 8.45]	<0.01
Neutrophils ($1×10^9$/L, ref: 1.8–6.3)	3.62 [2.59, 5.18]	4.78 [3.56, 6.34]	<0.01	3.89 [3.31, 5.20]	4.54 [3.66, 6.17]	<0.01
Erythrocytes ($1×10^{12}$/L, ref: 4.3–5.8)	4.34 [3.94, 4.79]	4.69 [4.46, 4.95]	<0.01	4.56[4.01, 5.24]	4.83 [4.21, 5.67]	<0.01
Hemoglobin (g/L, ref: 130–175)	136 [125, 148]	131 [124, 138]	<0.01	134 [128, 142]	136 [129, 140]	0.08
Platelet ($1×10^9$/L, ref: 125–350)	174 [133, 211]	219 [188, 258]	<0.01	182 [142, 231]	211 [179, 254]	<0.01
Monocytes ($1×10^9$/L, ref: 0.1–0.6)	0.46 [0.33, 0.62]	0.78 [0.61, 1.00]	<0.01	0.48 [0.36, 0.63]	0.59 [0.44, 0.74]	<0.01
Lymphocytes ($1×10^9$/L, ref: 1.1–3.2)	1.02 [0.72, 1.48]	1.27 [0.89, 1.76]	<0.01	1.21 [0.84, 1.74]	1.13 [0.83, 1.68]	0.49
Blood biochemistry						
AST (U/L, ref: 15.0–40.0)	42 [33, 56]	29 [20, 45]	0.02	47 [36, 64]	28 [22, 30]	0.01
ALT (U/L, ref: 9.0–50.0)	23 [15, 39]	14 [12, 20]	<0.01	24 [17, 33]	15 [13, 17]	0.01
BUN (mmol/L, ref: 2.8–7.6)	4.5 [3.5, 6.1]	4.0 [3.6, 5.1]	0.44	5.3 [4.4, 6.0[	4.4 [3.9, 4.7]	0.047
Creatinine (μmol/L, ref: 64.0–104.0)	70.6 [57.0, 85.5]	49.0 [37.5, 85.4]	0.04	68.4 [60.2, 71.0]	55.9 [53.2, 60.4]	0.04
hsTnI (pg/mL, ref: 0.0–26.2)	7.2 [3.6, 17.4]	1.2 [0.9, 3.5]	<0.01	6.8 [2.6, 15.2]	1.9 [0.7, 2.8]	0.01
LDH (U/L, ref: 125–243)	212 [176, 338]	214 [181, 282]	0.78	217 [188, 309]	208 [187, 265]	0.07
CK-MB (U/L, ref: 0–25)	13 [10, 18]	13 [2, 16]	0.34	12 [4, 17]	10 [5, 16]	0.41
Coagulation function						
APTT (sec, ref: 28.0–43.5)	30.9 [28.3, 33.0]	35.7 [31.3, 39.4]	0.02	33.1 [30.1, 34.2]	34.0 [30.9, 37.2]	0.05
PT (sec, ref: 11.0–16.0)	12.7 [11.9, 13.5]	14.8 [13.8, 16.1]	0.01	13.3 [12.0, 13.5]	14.2 [12.5, 15.4]	0.048
D-dimer (ng/mL, ref: 0.0–500.0)	273.0 [150.0, 691.0]	125.5 [99.3, 2606.5]	0.45	317.4 [189.2, 745.2]	142.7 [109.5, 543.7]	0.04
Infection markers						
hsCRP (mg/L, ref: 0.0–3.0)	14.8 [3.4, 44.9]	2.2 [0.50, 6.9]	<0.01	10.1 [3.1, 26.9]	5.0 [1.9, 5.0]	<0.01
Procalcitonin (ng/mL, ref: < 0.5)	0.05 [0.05, 0.16]	0.10 [0.06, 98.81]	0.14	0.12 [0.05, 0.18]	0.14 [0.07, 0.23]	0.17

Notes: Data are expressed as counts with percentage, otherwise median [IQR]. Data from ZHWU were used for model development.
Abbreviations: ALT, alanine aminotransferase; APTT, activated partial thromboplastin time; AST, aspartate transaminase; BUN, blood urea nitrogen; CK-MB, creatine kinase isoenzyme MB; DBP, diastolic blood pressure; hsCRP, high-sensitivity C-reactive protein; hsTnI, high-sensitivity Troponin I; LDH, lactate dehydrogenase; NT-proBNP, N-terminal pro b-type natriuretic peptide; PT, prothrombin time; RR, respiration rate; SBP, systolic blood pressure; SpO_2, pulse oxygen saturation.

Table 2 Performance of the Multi-Tree XGBoost in Identifying COVID-19 from Influenza Using 100-Round 5-Fold Cross-Validation

	AUC for Training Sets (95% CI)	AUC for Testing Sets (95% CI)
Age	0.92 (0.92, 0.92)	0.91 (0.91, 0.92)
Age + hsCRP	0.94 (0.94, 0.94)	0.91 (0.91, 0.92)
Age + hsCRP + Monocytes	0.96 (0.96, 0.97)	0.93 (0.93, 0.93)

Note: Data from ZHWU were used for model training and testing in the ratio 7:3.
Abbreviations: AUC, area under the receiver operating characteristic curve; hsCRP, high-sensitivity C-reactive protein.

performance of the model was compared with logistic regression. Our results indicated a superior performance of the proposed model as compared with the standard method (Table S4, Figure S6). Finally, the subgroup analysis demonstrated similar performance of the proposed model in identifying COVID-19 from influenza type A and type B, with an accuracy of 0.96 (AUC, 0.94 (95% CI 0.92, 0.96)) and 0.95 (AUC, 0.94 (95% CI 0.92, 0.96)), respectively (Table S5, S6 and Figure S7-S10).

The structure of an interpretable decision tree was obtained by a split of the 357 COVID-19 and 1893 influenza patients from ZHWU after data imputation (Figure 3). Our decision tree showed that cases with an old age (>16 years), a high hsCRP level (>14.2 mg/L) and a low cell count of monocytes ($\leq 0.68 \times 10^9$/L) are associated with a prediction favoring the diagnosis of COVID-19. However, 40 COVID-19 patients were incorrectly classified as influenza patients. Detailed analysis indicated that all those COVID-19 patients were non-severe, including 33 mild and 7 common cases. The three COVID-19 cases incorrectly classified by age showed normal levels of serum hsCRP and cell counts of circulating monocytes (Table S7).

Discussion

The major contribution of this study is providing a feasible and reliable decision tool to rapidly and
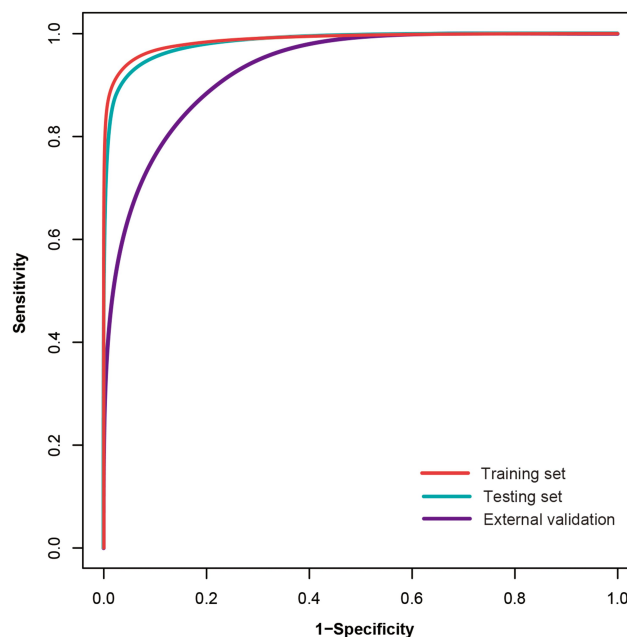

Figure 2 Receiver operating characteristic curves. This indicate the results for training, testing and external test sets. AUC: area under the receiver operating characteristic curve.

accurately distinguish between COVID-19 and influenza cases. The massive co-occurrences of COVID-19 and influenza cases may lead to relative shortage of detection kits and human resources, and even the collapse of the healthcare system. In addition, considering the current course of the COVID-19 pandemic, the ability of controlling the situation in the coming winter is an important and questionable issue worldwide. Under such circumstances, a rapid and accurate differential diagnosis of COVID-19 and influenza is the key step to initiate corresponding differential managements of suspected patients. Importantly, radiological examination and laboratory testing for specific pathogens are generally unavailable at the very first moments of hospital visit. Therefore, the availability of a reliable and feasible diagnosis tool based on very simple features may be of great importance in future daily practice. With such a perspective, the performance of the machine learning model reported in this study at

Table 3 Performance of the Single-Tree XGBoost Algorithm for COVID-19 Prediction

	Specificity (95% CI)	Sensitivity (95% CI)	NPV (95% CI)	PPV (95% CI)	Accuracy (95% CI)	AUC Score (95% CI)
Training set	0.98 (0.97, 0.99)	0.91 (0.87, 0.94)	0.98 (0.98, 0.99)	0.89 (0.86, 0.93)	0.97 (0.97, 0.97)	0.94 (0.93, 0.96)
Testing set	0.98 (0.96, 0.99)	0.88 (0.82, 0.94)	0.98 (0.96, 0.99)	0.87 (0.81, 0.94)	0.96 (0.96, 0.96)	0.93 (0.90, 0.96)
External test set	0.77 (0.73, 0.82)	0.91 (0.87, 0.94)	0.89 (0.86, 0.93)	0.80 (0.76, 0.84)	0.84 (0.84, 0.84)	0.84 (0.81, 0.87)

Abbreviations: NPV, negative prediction value; PPV, positive prediction value; AUC, area under the receiver operating characteristic curve.
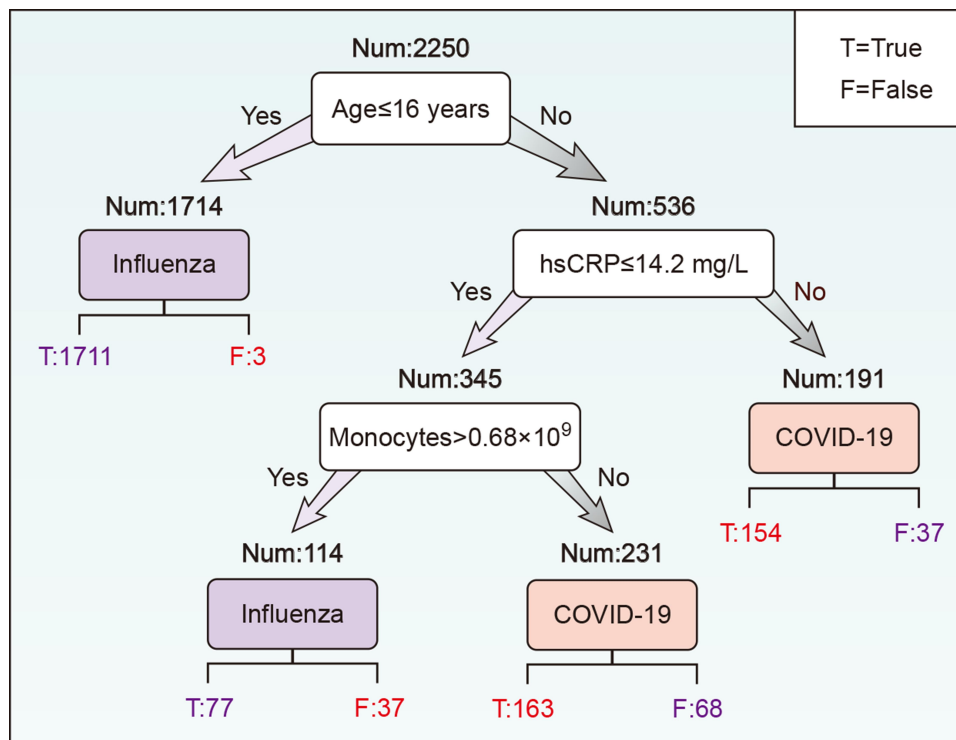
Figure 3 Clinically interpretable decision tree. Age, hsCRP, and monocyte were selected as key features for the decision tree. Finally, 357 COVID-19 and 1893 influenza patients were used for the development of the decision tree. hsCRP: high-sensitivity C-reactive protein, Num: number of patients.

least advocates that this proposed framework constitutes an attractive approach.

Our results suggested that age had the greatest impact on the predictions, with older ages driving the prediction towards COVID-19 and younger ages driving the predictions towards influenza. These findings are in line with the findings of the age distribution of COVID-19 and influenza patients. Current evidences indicate that COVID-19 is more likely to affect older adults.[10,11] Conversely, the seasonal influenza is commonly found among school-aged children, adolescents, and younger adults.[12–14] Although these studies demonstrate marked differences in the age distribution between COVID-19 and seasonal influenza, there are still quite a few young COVID-19 and old influenza patients.[15–17] Hence, it is difficult to classify these two diseases correctly when considering only the age of the patient. The proposed model suggested that it is difficult to make the prediction for those cases who had an age above 16 years unless involving hsCRP and monocytes.

Increased serum hsCRP is one of the clinical markers of a cytokine storm.[18] Hence, it is not surprising that hsCRP levels are elevated in virtually all COVID-19 patients[19] and serves as an important indicator of

a worsening outcome for COVID-19 patients.[9] In addition, a few studies have indicated that CRP could serve as a predictor of the illness severity for influenza A infection.[20] In the present study, the serum levels of hsCRP were markedly increased in COVID-19 patients as compared with influenza cases. Previous findings in COVID-19 and knowledge from the SARS-CoV-1 epidemic suggested that monocytes are possible participants in a cytokine storm and associated pathologies in COVID-19.[21–24] However, a very recent study reported that peripheral monocytes do not express substantial amounts of pro-inflammatory cytokines, suggesting that circulating monocytes do not significantly contribute to the cytokine storm in COVID-19.[25] In contrast, the influenza virus infection has been confirmed to be associated with a cytokine storm,[26] and monocytes are widely involved in the immune response to influenza virus infection.[27] Our results suggested that the cell counts of monocytes were significantly lower in COVID-19 cases than that of influenza ones. Finally, our prediction model selected hsCRP and monocytes as important predictors to distinguish between COVID-19 and influenza, and higher hsCRP or lower cell count of monocytes drives the prediction towards COVID-19.

The COVID-19 and influenza both cause respiratory disorders, which presents as a wide range of illnesses from asymptomatic or mild through to severe disease and death.[28] Hence, it is not surprising that clinical symptoms had no important influence on the COVID-19 prediction in the present study. Although reports have indicated that the loss of taste and smell could be common symptoms of COVID-19[29–31] and should be considered as a distinguishing symptom,[32] this has not been documented as a common symptom in the Chinese population.[11,33] Therefore, whether the loss of taste and smell could be key features for the classification of COVID-19 and influenza among the people in other countries is unclear. However, we still believe that these symptoms are not appropriate features for the development of a decision model, because a considerable number of COVID-19 cases are asymptomatic.[34–36]

This study has some limitations. First, this is a two-centered study, which provides primary assessment of potential features allowing to distinguish SARS-CoV-2 and influenza virus infections. The variability of the case-mix between centers will inherently lead to prediction models that will vary from one center to another, whenever the learning phase is conducted on a dataset from a single center and tested on cases from a different center. Therefore, the present model should be first considered as a pioneering tool proposing a promising version which performances will be likely increased when the algorithm will be fed with numerous data from many centers. Although external test resulted in a rather good assessment of the algorithm of generalizability, the two studying sites have similar hospital levels, similar geographic locations and same subtypes of SARS-CoV-2.[37] Therefore, the quality of model predictions in other hospitals, or considering populations out of China remains uncertain. Nevertheless, in the present two-center study, the classification results issued from the proposed machine learning-based model appear as very attractive. Second, the epidemiological and radiological data of patients have not been included for model development in the present study. Since the exposure history for the majority of COVID-19 patients is not definitive,[11] we have not considered such epidemiological data as potential indicators for prediction in the present study. However, since the purpose of our study was developing a rapid decision tool enabling a triage of COVID-19 and influenza cases, radiological examinations are not appropriate options because they are time-consuming. Third, because the proportion of other types of influenza is extremely low in our two studying sites, only those

patients with influenza virus type A and B infections were enrolled for analysis. Therefore, the potential quality of the prediction model in the presence of other types of influenza (including those that might emerge in the influenza epidemics of the next years) is inherently uncertain. However, our model showed no differences between identifying COVID-19 from influenza type A and B.

Conclusions

The proposed model selected age, hsCRP and monocytes as meaningful indicators for the classification of COVID-19 and influenza. The study demonstrates that a machine learning-based approach provides an attractive tool enabling a rapid and accurate triage tool for identifying COVID-19 from influenza. Such an approach would be particularly useful in regions having large number of COVID-19 and influenza cases while limited resources for laboratory testing of specific pathogens.

Abbreviation

AUC, area under the receiver operating characteristic curve; COVID-19, the novel coronavirus disease 2019; hsCRP, high-sensitivity C-reactive protein; NPV, negative predictive value; PPV, positive predictive value; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; WNH, Wuhan No.1 Hospital; ZHWU, Zhongnan Hospital of Wuhan University.

Data Sharing Statement

Additional information is available on request from the corresponding author (doctoryanzhao@whu.edu.cn).

Ethic Approval

This study was performed in accordance with the Declaration of Helsinki and approved by the Medical Ethics Committee, Zhongnan Hospital of Wuhan University (Clinical Ethical Approval No. 2,020,020). The ethics committee waived written informed consent because of the urgent need of data collection on COVID-19. All the data were deidentified to protect patient privacy.

Acknowledgments

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work. Xianlong Zhou and Zhichao Wang should be considered as co-first authors.

Funding

Disclosure

The authors report no conflicts of interest related to this work.

References

1. Cash R, Patel V. Has COVID-19 subverted global health? *Lancet*. 2020;395:1687–1688. doi:10.1016/S0140-6736(20)31089-8
2. World Health Organization. WHO Coronavirus Disease (COVID-19) dashboard. Available from: https://covid19.who.int/?gclid=Cj0KCQjwgo_5BRDuARIsADDEntTS0zycz7Wql-Ei1NsaK2oHhFzXLJZu5UtKk-t2xfmcXZXOZERtPMMaApHBEALw_wcB. Accessed October 1, 2020.
3. The Washington Post. CDC director warns second wave of corona-virus is likely to be even more devastating. Published April 22, 2020. https://www.washingtonpost.com/health/2020/04/21/coronavirus-secondwave-cdcdirector/.
4. Heymann DL, Shindo N. Scientific WHO, technical advisory group for infectious H. COVID-19: what is next for public health? *Lancet*. 2020;395:542–545. doi:10.1016/S0140-6736(20)30374-3
5. Zhou X, Ding G, Shu T, et al. The outbreak of coronavirus disease 2019 interfered with influenza in Wuhan. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3555239. Accessed March 25, 2020.
6. World Health Organization. Coronavirus disease (COVID-19) technical guidance: laboratory testing for 2019-nCoV in humans; Published 2020. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-guidance. Accessed January 28, 2021.
7. Chinese Natinal Influenza Center. Weekly china influenza surveillance report (Issue 8, 2020). Available from: http://ivdc.chinacdc.cn/cnic/zyzx/lgzb/202002/P020200229474997456548.pdf. Accessed February 29, 2020.
8. Chen T, Guestrin C Xgboost: a scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. New York; NY; United States. 2016: 785–794.
9. Yan L, Zhang HT, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*. 2020;1–6.
10. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020;382:1708–1720. doi:10.1056/NEJMoa2002032
11. Richardson S, Hirsch JS, Narasimhan M, et al. Presenting character-istics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area. *JAMA*. 2020; 323:2052–2059. doi:10.1001/jama.2020.6775
12. Karageorgopoulos DE, Vouloumanou EK, Korbila IP, Kapaskelis A, Falagas ME. Age distribution of cases of 2009 (H1N1) pandemic influenza in comparison with seasonal influenza. *PLoS One*. 2011;6: e21690. doi:10.1371/journal.pone.0021690
13. Khiabanian H, Farrell GM, St George K, Rabadan R. Differences in patient age distribution between influenza A subtypes. *PLoS One*. 2009;4:e6832. doi:10.1371/journal.pone.0006832
14. Wang XL, Yang L, Chan KH, et al. Age and sex differences in rates of influenza-associated hospitalizations in Hong Kong. *Am J Epidemiol*. 2015;182:335–344. doi:10.1093/aje/kwv068
15. Dong Y, Mo X, Hu Y, et al. Epidemiology of COVID-19 among children in China. *Pediatrics*. 2020;145:e20200702. doi:10.1542/peds.2020-0702
16. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395:497–506. doi:10.1016/S0140-6736(20)30183-5
17. Rüttimann RW, Bonvehí PE, Vilar-Compte D, Isturiz RE, Labarca JA, Vidal EI. Influenza among the elderly in the Americas: a consensus statement. *Rev Panam Salud Publica*. 2013;33:446–452.
18. Zhang W, Zhao Y, Zhang F, et al. The use of anti-inflammatory drugs in the treatment of people with severe coronavirus disease 2019 (COVID-19): the Perspectives of clinical immunologists from China. *Clin Immunol*. 2020;214:108393. doi:10.1016/j.clim.2020.108393
19. Han R, Huang L, Jiang H, Dong J, Peng H, Zhang D. Early clinical and CT manifestations of Coronavirus Disease 2019 (COVID-19) Pneumonia. *AJR Am J Roentgenol*. 2020;215:338–343. doi:10.2214/AJR.20.22961
20. Zimmerman O, Rogowski O, Aviram G, et al. C-reactive protein serum levels as an early predictor of outcome in patients with pan-demic H1N1 influenza A virus infection. *BMC Infect Dis*. 2010;10:288. doi:10.1186/1471-2334-10-288
21. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol*. 2020;20:363–374. doi:10.1038/s41577-020-0311-8
22. Zhou Z, Ren L, Zhang L, et al. Heightened innate immune responses in the respiratory tract of COVID-19 patients. *Cell Host Microbe*. 2020;27(6):883–890.e2. doi:10.1016/j.chom.2020.04.017
23. Giamarellos-Bourboulis EJ, Netea MG, Rovina N, et al. Complex immune dysregulation in COVID-19 patients with severe respiratory failure. *Cell Host Microbe*. 2020;27:992–1000.e3. doi:10.1016/j.chom.2020.04.009
24. Yip MS, Leung NH, Cheung CY, et al. Antibody-dependent infection of human macrophages by severe acute respiratory syndrome coronavirus. *Virol J*. 2014;11:82. doi:10.1186/1743-422X-11-82
25. Wilk AJ, Rustagi A, Zhao NQ, et al. A single-cell atlas of the peripheral immune response to severe COVID-19. Preprint. *medRxiv. 2020*. doi:10.1101/2020.04.17.20069930
26. Teijaro JR, Walsh KB, Rice S, Rosen H, Oldstone MB. Mapping the innate signaling cascade essential for cytokine storm during influenza virus infection. *Proc Natl Acad Sci U S A*. 2014;111:3799–3804. doi:10.1073/pnas.1400593111

27. Lamichhane PP, Samarasinghe AE. The role of innate leukocytes during influenza virus infection. *J Immunol Res*. 2019;2019: 8028725. doi:10.1155/2019/8028725

28. World Health Organization. Influenza and COVID-19 - similarities and differences. Available from: https://www.who.int/emergencies/ diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-de tail/q-a-similarities-and-differences-covid-19-and-influenza?gclid= Cj0KCQjwpZT5BRCdARIsAGEX0zkcbRzyHpdwwlOPDNoDOkP l71XZ9-8uKvkCzZ6mb8_yE3b3UJFZnYwaAo5KEALw_wcB. Accessed March 17, 2020.

29. Gautier JF, Ravussin Y. A new symptom of COVID-19: loss of taste and smell. *Obesity*. 2020;28:848. doi:10.1002/oby.22809

30. Mullol J, Alobid I, Mariño-Sánchez F, et al. The loss of smell and taste in the COVID-19 outbreak: a tale of many countries. *Curr Allergy Asthma Rep*. 2020;20:61. doi:10.1007/s11882-020-00961-1

31. Lechien JR, Chiesa-Estomba CM, De Siati DR, et al. Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study. *Eur Arch Otorhinolaryngol*. 2020;277:2251–2261. doi:10.1007/s00405-020-05965-1

32. Dawson P, Rabold EM, Laws RL, et al. Loss of taste and smell as distinguishing symptoms of COVID-19. *Clin Infect Dis*. 2020: ciaa799. doi: 10.1093/cid/ciaa799.

33. Mao L, Wang M, Chen S, et al. Neurological manifestations of hospitalized patients with COVID-19 in Wuhan, China: a retrospective case series study. *MedR*. 2020:xiv. doi: 10.1101/ 2020.02.22.20026500.

34. Tian S, Hu N, Lou J, et al. Characteristics of COVID-19 infection in Beijing. *J Infect*. 2020;80:401–406. doi:10.1016/j.jinf.2020.02.018

35. Kim GU, Kim MJ, Ra SH, et al. Clinical characteristics of asymptomatic and symptomatic patients with mild COVID-19. *Clin Microbiol Infect*. 2020;26:948.e1–948.e3. doi:10.1016/j.cmi.2020.04.040

36. Zhu J, Ji P, Pang J, et al. Clinical characteristics of 3062 COVID-19 patients: a meta-analysis. *J Med Virol*. 2020. doi:10.1002/jmv.25884

37. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A*. 2020;117:9241–9243. doi:10.1073/pnas.2004999117