European Journal of Radiology Open 13 (2024) 100603

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Original article

A tri-light warning system for hospitalized COVID-19 patients: Credibility-based risk stratification for future pandemic preparedness

Chuanjun Xu [a,1], Qinmei Xu [b,1], Li Liu [c,1], Mu Zhou [d], Zijian Xing [e], Zhen Zhou [e], Danyang Ren [f], Changsheng Zhou [g], Longjiang Zhang [g], Xiao Li [g,*], Xianghao Zhan [h,*], Olivier Gevaert [b,*], Guangming Lu [g,*]

[a] Department of Radiology, the Second Hospital of Nanjing, Nanjing University of Chinese Medicine, Nanjing 210003, China
[b] Department of Biomedical Data Science (BMIR), Department of Medicine, Stanford University, Stanford, CA 94304, USA
[c] Department of Computer Science, University of California Santa Cruz, Santa Cruze, CA 95064, USA
[d] Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA
[e] Department of Deepwise AI Lab, Deepwise Inc., Beijing, China
[f] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[g] Department of Medical Imaging, Jinling Hospital, Nanjing, Jiangsu, China
[h] Department of Bioengineering, Stanford University, Stanford 94305, USA

ARTICLE INFO

Keywords:
COVID-19 pandemic
Multi-modal artificial intelligence
Risk stratification
Conformal prediction
Multi-center study

ABSTRACT

Purpose: The novel coronavirus pneumonia (COVID-19) has continually spread and mutated, requiring a patient risk stratification system to optimize medical resources and improve pandemic response. We aimed to develop a conformal prediction-based tri-light warning system for stratifying COVID-19 patients, applicable to both original and emerging variants.

Methods: We retrospectively collected data from 3646 patients across multiple centers in China. The dataset was divided into a training set (n = 1451), a validation set (n = 662), an external test set from Huoshenshan Field Hospital (n = 1263), and a specific test set for Delta and Omicron variants (n = 544). The tri-light warning system extracts radiomic features from CT (computed tomography) and integrates clinical records to classify patients into high-risk (red), uncertain-risk (yellow), and low-risk (green) categories. Models were built to predict ICU (intensive care unit) admissions (adverse cases in training/validation/Huoshenshan/variant test sets: n = 39/21/262/11) and were evaluated using AUROC ((area under the receiver operating characteristic curve)) and AUPRC ((area under the precision-recall curve)) metrics.

Results: The dataset included 1830 men (50.2 %) and 1816 women (50.8 %), with a median age of 53.7 years (IQR [interquartile range]: 42–65 years). The system demonstrated strong performance under data distribution shifts, with AUROC of 0.89 and AUPRC of 0.42 for original strains, and AUROC of 0.77–0.85 and AUPRC of 0.51–0.60 for variants.

Conclusion: The tri-light warning system can enhance pandemic responses by effectively stratifying COVID-19 patients under varying conditions and data shifts.

1. Introduction

Coronavirus disease 2019 (COVID-19) continues to spread and has caused over 775 million confirmed cases and over seven million deaths by 28 April 2024 [1]. Several mutant strains emerged with increased infectiousness (e.g. Omicron) or morbidity (e.g. Delta) when compared with the previously observed strain in the pandemic[2]. These have been designated as "variants of concern" by the WHO.[3]. "Variants of concern" can suddenly emerge and spread, leading to near-capacity usage of hospitals and intensive care units.[4]. Therefore, reliable, generalizable, and sustainable methods for the timely identification of high-risk patients are crucial for clinical decision-making and efficient

* Corresponding authors.
E-mail addresses: xqm@smail.nju.edu.cn (X. Li), xzhan96@stanford.edu (X. Zhan), ogevaert@stanford.edu (O. Gevaert), cjr.luguangming@vip.163.com (G. Lu).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.ejro.2024.100603
Received 30 July 2024; Received in revised form 12 September 2024; Accepted 30 September 2024

allocation of resources in the context of existing and emerging virus strains.

COVID-19 is primarily characterized by pulmonary inflammatory lesions, where computed tomography (CT) feature assessment by radiologists is used for treatment evaluation [5]. However, current processes are often subjective and unable to accurately predict the disease progression, leading to a significant increase in the workload for radiologists. Therefore, an automated and quantitative analytical method of analyzing CT images is urgently needed to provide more objective and reliable evaluation for better determination of disease progression.

Recent studies have confirmed that an image-based AI prognostic prediction model can play a supportive role in the diagnosis and prognosis prediction of COVID-19[3,6,7]. For instance, AI-based models utilizing CT radiomic features and clinical indicators can predict high-risk events during hospitalization and events such as in-hospital mortality, which demonstrated the value of imaging and clinical features in prognostic prediction[7]. Clinical information such as sex, age, symptoms, comorbidities, and laboratory values of patients are expected to be used for a more accurate assessment of prognosis [8,9]. However, current research efforts do not address the generalizability of models under data distributions caused by mutant strains and variability related to the hospital setting. Because of the changing status of virus mutation, models developed based on existing strains may not be applicable to new strains. In addition, unlike evidence gained from small cohorts[10,11], we emphasize the necessity of multi-center evaluation for testing model generalization. Clinical centers are often equipped with various medical devices, protocols, and resources. For instance, field hospitals and permanent hospitals for COVID-19 patients can have different conditions of equipment, medical personnel, and protocols. Such differences caused by the hospital types increase the difficulty of measuring model performance. Further, existing studies lack the analysis of prediction reliability, i.e., how much confidence we have for a particular prediction and is high prediction confidence related to high prediction accuracy.

Conformal prediction is a computational framework that promises to effectively quantify prediction uncertainty [12]. Based on the weak assumption of independently and identically distributed data, it leverages the empirical distribution of the nonconformity measurement of the training data to calibrate the uncertainty of a particular prediction. In the biomedical field, conformal prediction has been used in uncertainty quantification for breath-air-based lung cancer prediction [13], large-scale whole-slide tissue images classification [14], alternative medicine discrimination [15], biomedical natural language processing for fast literature filtering [16] and RNA-sequencing-based breast cancer subtyping [17]. These applications manifest the effectiveness of conformal prediction in uncertainty quantification in biomedical applications.

To address these challenges, we collected PCR-confirmed COVID-19 patients from 40 hospitals across China, developing an end-to-end flexible tri-light warning system to predict the potential requirement of ICU care of patients (Fig. 1). This system is based on conformal prediction that calculates the credibility of each prediction. Briefly, this system stratifies patients into three categories: a. Red: high probability with high credibility, represents high-risk patients that need ICU care within 28 days; b. Green: low probability with high credibility, represents low-risk patients; c. Yellow: high/low probability with low credibility, represents patients with uncertainty of risk that need further monitoring.

In this study, we hypothesize that conformal prediction enables effective quantification of the prediction uncertainty in COVID-19 patient ICU admission prediction, aids flexible patient stratification based on the prediction uncertainty and generalizes well across data collected from different medical centers across COVID-19 variants. The major contributions of this study can be summarized as follows: first, based on conformal prediction, we developed an early warning system for hospitalized COVID-19 patients. This system quantifies the credibility of each prediction when calculating the probability of progression, providing important clinical guidance for patient management and treatment mode selection. Second, we proposed a flexible tri-light warning strategy, which can change the proportion of high-risk patients from model output according to local medical resource allocation and virulence of the virus, helping to optimize the allocation of medical resources, thus achieving closer monitoring and timely treatment for high-risk patients. Third, we collect a multi-center cohort from 40 hospitals in China (n = 8721) for systematic evaluation that emphasizes the

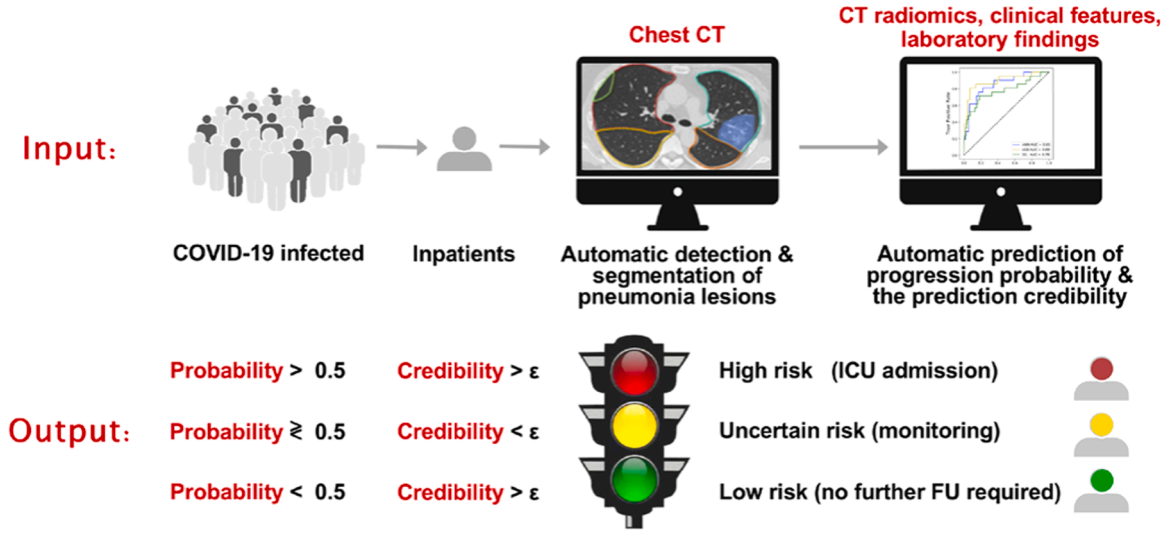

Fig. 1. Workflow of the proposed Tri-light Warning System for hospitalized COVID-19 patients. First, the system automatically detects and segments the pneumonia lesions on the input chest CT images of patients at admission, and extracts the radiomic features from the lesion. Second, the radiomic features, clinical features, and laboratory findings of patients were combined to automatically predict whether the patient's condition will progress and require ICU care within 28 days, and the prediction credibility was also given. Finally, the system classifies patients into high-risk, low-risk, and uncertain risks according to the progression probability and the prediction credibility, aiming to provide closer medical monitoring for high-risk and uncertain patients and dispatch medical resources in advance to ensure the timely treatment of high-risk patients. Particularly, medical institutions can adjust the credibility threshold ϵ of the system according to the virus characteristics (mortality, infectivity, etc.) and local medical resources, so as to flexibly change the proportion of high-risk patients output by the system and the reliability requirement of prediction. CT = computed tomography; ICU = intensive care unit; FU = follow-up.

test of model generalizability under data distribution caused by different types of hospitals and different variant strains. To assess the mutational effect on the virus, we perform analysis based on COVID-19 patients infected with the original strain and validate the performance on data from other hospitals with different strains (i.e. delta and omicron, n = 544) to evaluate the generalization of the model.

The remaining sections of this paper are organized as follows. Section 2 discusses related work. The architecture and implementation detail of the proposed method are presented in Section 3, followed by the experimental results in Section 4. Finally, the discussion and concluding remarks are given in Section 5.

2. BACKGROUND: AI-enabled COVID-19 studies

Predicting patient outcomes with COVID-19 at an early stage is crucial to optimize clinical care and medical resource management [18]. Multiple AI models based on machine learning (ML) and deep learning have been proposed to address this task. Example models estimated mortality risk in patients with suspected or confirmed COVID-19 [19–22]. Other models aimed to predict progression to a severe or critical state [10,23]. There are also efforts to predict the length of hospital stay [9,24]. The most common prognostic predictors included age [9,25,26], sex [27–29], comorbidity (including hypertension, diabetes, cardiovascular disease, and respiratory disease) [9,30], lymphocyte count [26,31], and also radiomic features derived from CT images [32–34].

Recently developed models share a similar perspective with our research and show the potential value of the clinical application.A study utilized CT-based radiomic features and machine learning algorithms to accurately predict the stages of COVID-19 infection, including normal, mild, moderate, and severe stages, with accuracies of 99.12 %, 98.24 %, 98.73 %, and 99.9 %, respectively.[35]. Another study developed a CT feature-based predictive model using deep neural networks (DNN) to identify asymptomatic carriers, achieving an AUC of 0.898[36]. Also, our previous study used 3522 PCR-confirmed COVID-19 inpatients from 39 hospitals and performed CT-based analysis combined with electronic health records and clinical laboratory results with prognostic estimation for the rapid risk stratification (AUROC 0.916–0.919) [7].

However, in AI-based applications, the reliability of predictions is significant for assisted decision and risk control in real-world applications: unreliable prediction can interfere with the clinicians' decision and may lead to misdiagnosis putting huge pressure on the patients' families. Therefore, knowing how much confidence is associated with a prediction made by the model is important for the decision-making process for clinicians [37]. PROBAST analysis indicates that the majority of proposed models above are at a high risk of bias, and their reported performance is probably optimistic [32,38–40]. Unreliable predictions could cause more harm than benefit in guiding clinical decisions. Therefore, current AI-enabled models have not been validated or implemented outside of their original study sites, which are therefore not recommended in clinical practice[38,39,41]. To address this challenge, we developed prognostic prediction models based on a large, heterogeneous, real-world data set in a newly proposed conformal prediction framework. The models not only provide high prediction performance in predicting whether a patient in the general ward will be admitted to the intensive care units (ICUs) under distribution drifts but also provide the users with prediction reliability information. Based on the prediction reliability information given by the conformal prediction framework, a tri-light warning system is introduced to enable the users to adapt health policies for ICU resource allocation Fig. 1. This framework may also have potential applications in the SARS-CoV-2 omicron XBB.1.16 variant, which will soon spread globally[42].

3. Methods

3.1. Patient cohort

The data in this study were collected from 40 hospitals in China (n = 8721). Patients selection followed the inclusion criteria: (a) RT-PCR confirmed positive severe acute respiratory syndrome coronavirus (SARS-CoV-2) nucleic acid test; (b) baseline chest CT examinations and laboratory tests on admission; (c) short-term prognosis information (discharge or admission to ICU). Along with the exclusion criteria, we collected 3920 patients for analysis, including four cohorts. First, a training cohort (n = 1451) is set for model development, which included patients from 17 hospitals. Second, we performed model parameter tuning on a validation set (n = 662) which consisted of patients from nine independent medical centers. Third, we assessed the performance of models on an external test set (n = 1263) with a higher rate of ICU admission from Huoshenshan (HSS) field Hospital, an emergency specialty field hospital designed to treat people with COVID-19 in Wuhan, Hubei, China. In addition, we built a specific test set (n = 544) based on COVID-19 Delta and Omicron variants to evaluate the generalization of models. An overview of the patient cohorts is summarized in Fig. 2.

3.2. Data collection and preprocessing

Our multi-modal data for each patient included:

1) Clinical data based on electronic Health Records (EHR): (a) demographics: age and gender; (b) comorbidities: coronary heart disease, diabetes, hypertension, chronic obstructive lung disease (COPD), chronic liver disease, chronic kidney disease, and carcinoma; and (c) clinical symptoms: fever, cough, myalgia, fatigue, headache, nausea or vomiting, diarrhea, abdominal pain, and dyspnea on admission. To extract the clinical data from the free-text EHR in Chinese, we developed a rule-based language processing algorithm. Firstly, the clinical descriptions are segmented from EHR by splitting the paragraphs with subtitles. Then, we established a keyword list containing all the descriptions associated with a particular clinical feature, such as 'fever' and 'cough'. Simply using a regular expression to match keywords is not practical. An example of the clinical description is: "This patient had fever and cough three days ago, and he had no diarrhea or vomiting, and today he is transmitted to this hospital without fever". Since the symptoms can progress/recover, the same keywords have different meanings with negation and thought groups. Therefore, we designed a voting rule to extract the clinical data: by breaking down the clinical descriptions with commas, the keywords can be matched into several thought groups. In each thought group, the frequency of target keywords is recorded, with a default value of zero. If a thought group begins with negation (such as 'no', or 'did not find'), the keywords appeared to vote zero to their encoding values; Otherwise, the keywords appeared to vote one. The votes are summed up and compared with zero, and then the sum will convert into a boolean value and be viewed as the encoded value for a specific clinical feature. With the voting rule, the negative appearance of a keyword can be distinguished, and the sum of votes can provide more information for further study as it indicates the frequency of a symptom.
2) Laboratory test: blood routine, coagulation function, blood biochemistry, infection-related biomarkers. (a) blood routine: white blood cell (WBC) count ($\times 10^9/L$), neutrophil count ($\times 10^9/L$), lymphocyte count ($\times 10^9/L$), platelet count ($\times 10^9/L$), and hemoglobin (g/L); (b) coagulation function: prothrombin time (PT) (s), activated partial thromboplastin time (aPTT) (s), and D-dimer (mg/L); (c) blood biochemistry: albumin (g/L), alanine aminotransferase (ALT) (U/L), aspartate Aminotransferase (AST) (U/L), total bilirubin ($mmol/L$), serum potassium ($mmol/L$), sodium ($mmol/L$), creatinine ($\mu mol/L$), creatine kinase (CK) (U/L), lactate

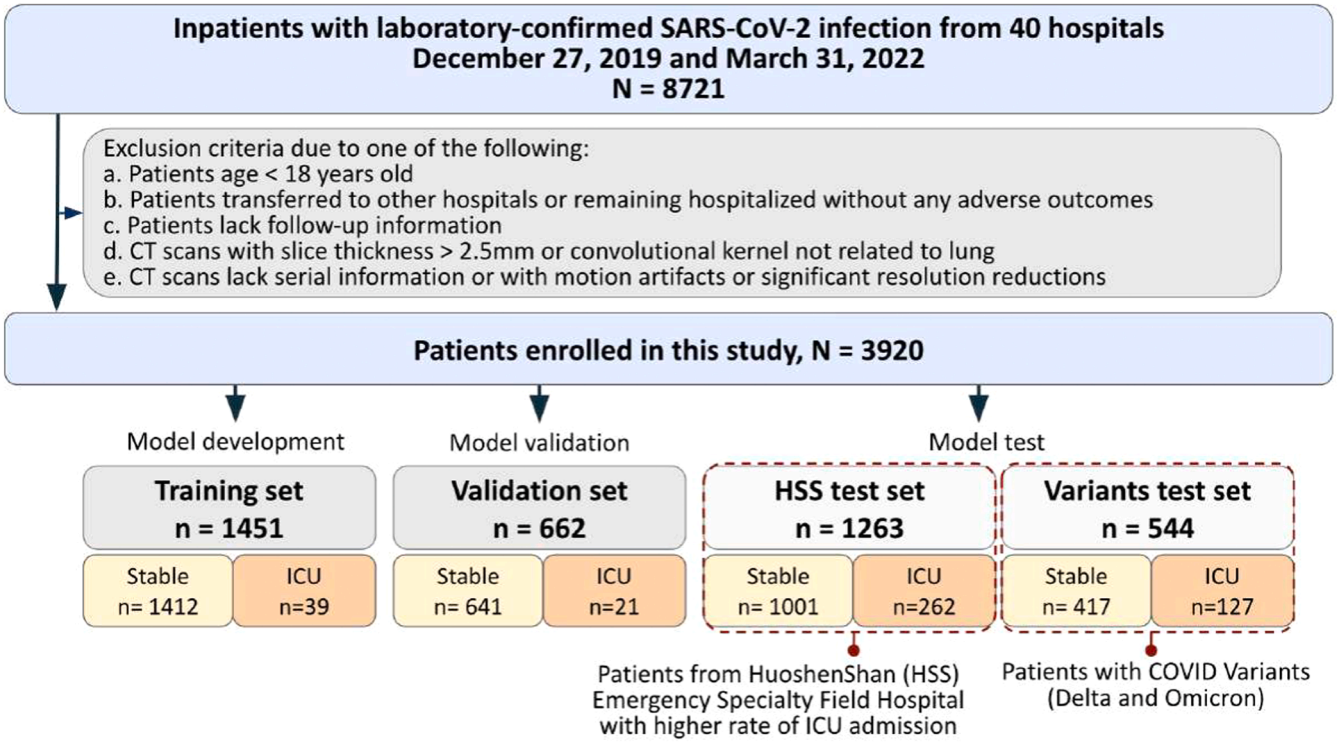

Fig. 2. The summary of the patient cohorts and exclusion criteria. The models were trained on the training set and the model hyperparameters were tuned on the validation set. Upon hyperparameter tuning and optimizing the feature selection approach on the validation set, the training set and validation set are combined to train the ultimate models. The ultimate models are tested on the two test cohorts under data distribution drifts: the Huoshenshan field hospital test set and the Delta and Omicron variant data set.

dehydrogenase (LDH) (U/L), α-Hydroxybutyrate dehydrogenase (HBDH) (U/L); (d) infection-related biomarkers: C-reactive protein (CRP) (mg/L). Patients took laboratory tests on the date of admission in the training set, validation set, and Variants test set, while patients in the HSS test set received laboratory tests within two days after admission due to the centralized outbreak in Wuhan and the limited medical resources. To alleviate missing values that occurred in records, we applied median imputation on the lab data when a missing rate was $\leq$50 %, which has been validated effective in the previous study [7]. Each inpatient received laboratory tests within 48 h after admission and only clinical data on or prior to the date of the CT were used for prediction.

3) CT radiomics: Patients took baseline CT scans within three days after admission [43]. It should be mentioned that Chinese guidelines for the diagnosis and treatment of novel coronavirus infection recommend the CT imaging features as one of the criteria for clinical classification, the treatment chosen, and the discharge criteria for hospitalized patients. Chest CT scans were performed using $\geq$16 slice multidetector CT scanners (Aquilion ONE / Aquilion PRIME / BrightSpeed / BrightSpeed S / Brilliance 16 / Brilliance 64 / Discovery CT750 HD / eCT / Fluorospot Compact FD / HiSpeed Dual / iCT 256 / Ingenuity CT / Ingenuity Flex / LightSpeed VCT / LightSpeed 16 / NeuViz 16 Classic / Optima CT520 Series / Optima CT540 / Optima CT680 Series / ScintCare CT 16E / Sensation 64 / SOMATOM Definition AS+ / SOMATOM Definition Flash / uCT 510) without use of iodinated contrast agents. To minimize motion artifacts, patients were asked to hold their breath, then axial CT images were acquired during end-inspiration. The CT scan protocols were as follows: tube voltage, 100–120 kVp; effective tube current, 110–250 mAs; detector collimation, 16–320 $\times$ 0.625–2.5 mm; slice thickness, 0.625–2.5 mm; pitch, 0.8–1.375. The CT images were reconstructed by iterative reconstruction technique if possible based on the raw data. A commercial deep-learning AI system (Beijing Deepwise & League of PhD Technology Co. Ltd) was first used to detect and segment the pneumonia lesion, and two radiologists (Q.M.X. and C.S.Z.) checked the results of the automatic segmentation. Then, pyradiomics (v3.0) running in the Linux platform was adopted to extract radiomic features (1652 features per lesion). Next, for a given patient and for each radiomic feature, we summarized the distribution of the feature values across all the lesions for the patient by several summary statistics (mean, median, standard deviation, skewness, the first quartile, the third quartile) and the number of lesions. Finally, a total of 9913 quantitative radiomic features were extracted from CT images for each patient.

3.3. Feature engineering

To address the imbalance in the data set and high dimensionality of the feature space before modeling, several different feature engineering approaches were applied to select/weigh the features and augment the minority cases:

1) synthetic minority oversampling technique (SMOTE): SMOTE oversamples the minority class by synthesizing new minority data. Under the assumption that data close in the feature space are similar in their labels, SMOTE randomly selects a pair of minority-class data, draws a line between them in the feature space, and finds a random point along the line segment as the new synthetic minority-class data. SMOTE was implemented with the Python package imblearn (version: 0.6.2);
2) feature selection based on shrunken centroids (SC): SC is an algorithm derived from the nearest centroids (NC) [44]. However, SC further attenuates the noisy features which do not have much class-related information. The SC feature selection works with the following steps (assume the original feature space has a dimensionality of D):

3) Initially, in the original feature space, we calculate the the centroids $\overline{x}_t \in \mathbf{R}^D$ of each class (1, 2, …, T) and the universal centroid μ of all samples are computed (Z: $((x_1, y_1), \ldots, (x_n, y_n))$). Here C_t refers to the sample set with label t, and n_t refers to the quantity of samples with label t.

$$\overline{x}_t = \sum_{j=1}^{C_t} \frac{x_{jt}}{n_t} \tag{1}$$

$$\mu = \sum_{j=1}^{n} \frac{x_j}{n} \tag{2}$$

Next, we compute the accumulated within-class standard deviation as a non-biased approximation of the standard deviation for the overall data distribution. Then, the differences between class centroids and the universal centroid are normalized using this within-class standard deviation:

$$\sigma^2 = \frac{1}{n-T} \sum_t \sum_{j \in Ct} \left(x_j - \overline{x}_t\right)^2 \tag{3}$$

$$d_t = (\overline{x}_t - \mu)/\sigma \tag{4}$$

Lastly, these differences are subject to shrinkage via a threshold symbolized by Δ, which is hyperparameter:

$$\begin{aligned} d_t' &= \operatorname{sign}(\mathrm{d}_t)(|\mathrm{d}_t| - \Delta)_+ \\ f_+ &= \begin{cases} f & f > 0 \\ 0 & f \le 0 \end{cases} \end{aligned} \tag{5}$$

The feature selection effect is governed by the threshold Δ: if d_{lm}, the difference of an l-th attribute of class m, has an absolute value less than the threshold Δ, the corresponding feature is deemed insufficiently informative for categorization. As a result, the difference in this attribute will be shrunken to zero, effectively eliminating the non-informative attribute and reducing the dimensionality of the data.

4) feature selection based on Lasso (Lasso): a logistic regression model is fitted with L1 penalty. By adding different strengths of L1 penalty, different numbers of features will be given zero coefficients in the logistic regression and the features with non-zero coefficients will be selected to reduce the dimensionality of the feature space;
5) feature weighting based on principal component analysis (PCA): PCA performs the covariance analysis and finds the principal components that maximize the variance of the data projections. By projecting the features onto the principal components, the dimensionality of the original feature space is reduced and the information is compressed in the projections on the principal components. It should be noted that generally, PCA only weighs the features but does not mask features or perform feature selection. Lasso and PCA were implemented with the scikit-learn package (version: 0.21.3).

In this study, SMOTE was first implemented to cope with the class imbalance and then one of the feature selection/weighing methods (SC/Lasso/PCA) was used. The hyperparameters associated with these algorithms that have been tuned in this study include the threshold Δ for SC feature selection, the strength of the L1 penalty for Lasso feature selection C, and the number of principal components selected for PCA.

3.4. Prediction model development and evaluation

In this study, the task is to predict whether a patient admitted to the general ward will be admitted to the intensive care unit (ICU) within 28 days. To address this classification task, we concatenated the CT radiomic features, clinical features (including the demographics, clinical symptoms, and comorbidities), and lab test features. Then, we leveraged the shrunken centroids (SC)[44,45], Light Gradient Boosting Machine (LGB)[46] and artificial neural network (ANN) [47]. These algorithms are chosen because they are the representatives of different classification rationales: SC classifies samples based on the similarity in the Euclidean distance in the feature space (as a modified version of the nearest centroids algorithm) [44,45], LGB is a tree-based ensemble-learning algorithm with the boosting ensemble-learning strategy [46], and ANN is a representative of the deep learning technology based on the gradient descent in minimizing the binary cross-entropy loss function [47]. The SC was implemented with Python 3.7 [45]. The LGB was implemented with the Python package lightgbm. The ANN was implemented with the Python package scikit-learn (version: 0.21.3).

The hyperparameters are tuned based on the performance on the validation set in this study and the types of hyperparameters tuned are listed as follows: 1) for SC: the threshold Δ; 2) for LGB: the learning rate, the maximum depth of trees and the number of leaves of the trees; 3) for ANN: the number of hidden layers and the numbers of hidden units for each hidden layer. It should be mentioned that the classifier hyperparameters are tuned in combination with the feature selection methods and feature selection hyperparameters on the validation data set.

3.5. Analysis of important features in decision

To provide users with more explainable and interpretable decision-making based on the model, we investigated the important features that support the decision-making process based on the simplest and most interpretable models investigated in this study: SC. The importance is measured by the absolute value of the shrunken contrasts d_t'. After the soft-thresholding operation, the irrelevant and noisy features will be set to zero, and the other features will be reduced by the threshold Δ. Therefore, the valuable features will remain high in the vector d_t', and their absolute values directly indicate the weights in the prediction process. To ensure the robust important features are investigated, with 30 times of bootstrapping, we calculated the mean absolute values of shrunken contrast for each feature on the shrunken centroid for class 'positive'. The higher mean absolute value of the shrunken contrasts indicates that the corresponding feature is given more weight by the model. Meanwhile, the signs of the shrunken contrast for the features indicate the positive/negative association with a positive prediction: i.e., for the j-th feature, if d_{+j}' is positive, a higher feature value contributes to a positive prediction, while if d_{+j}' is negative, a lower feature value contributes to a positive prediction. Additionally, we also investigated the relative feature importance of lab test data, clinical data, and radiomics data.

3.6. Reliability quantification with conformal prediction

In the application of predicting ICU admission of COVID-19 patients, besides the predictions themselves, the prediction reliability and uncertainty are of great importance in the management of health policies to optimize the usage of the limited ICU resources. Once the prediction reliability can be quantified, the ICU utilization policies can be made according to both how confident one has for a positive prediction and how severe the outcomes can be brought about by the current virus variant.

To quantify the prediction reliability, we leveraged the conformal prediction which was developed by Vladimir Vovk [12]. Conformal prediction assumes that the data abide by the independent and identical

distribution (I.I.D) and outputs credibility as the reliability information for each prediction. The applications and brief introduction of conformal predictors can be found in previous publications [15,45,48]. To briefly introduce the conformal prediction framework, firstly, the nonconformity measurements are computed based on heuristic rules/algorithms such as based on the conditional probability output by the classification model (e.g. conformal prediction with shrunken centroids (CPSC) [45] or based on the ratio of the cumulative distance of the heterogeneous samples and homogeneous samples (e.g., conformal prediction with k-nearest neighbors [13,15]). Then, the nonconformity measurements of the training data and the test data are computed and the nonconformity measurement is then used as a statistic to be calibrated. For the training data, the nonconformity measurements are computed for the combination of the feature and ground truth label, while on the test data, the nonconformity measurements are computed for all possible labels in the label space. The conformal prediction then takes the empirical distribution of the nonconformity measurements on the training dataset as the true distribution and evaluates which percentile the nonconformity measurement of the test sample-label combination lies. Based on the percentile, the conformal prediction computes the p-value reflecting the conformity to the training data distribution of a particular feature-label combination, i.e., how well the test feature-label combination conforms to the training data distribution and quantifies the prediction credibility based on the most probable label's conformity to the training data distribution. The detailed methods are listed below:

To compute the credibility, we leverage the conformal predictor based on the prediction probability, and the steps can be summarized as the following steps:

1) Convert the predicted probability to a nonconformity measurement: a metric to quantify how well a particular feature-label combination conforms to the training data. Here, we leveraged a design of the nonconformity measurement α_i that has been validated in multiple machine learning applications [37, 49]:

$$\alpha_i = 0.5 - \frac{\widehat{p}(y_i|x_i) - \max\widehat{p}_{y!=y_i}(y_i|x_i)}{2} \tag{6}$$

Here y_i and x_i denote the label and feature of the $i-th$ sample. In this study, the predicted probability can be computed by SC, LGB, or MLP.

2) Based on the nonconformity measurement, all the training samples' nonconformity measurement values can be computed and the distribution will be further used to calibrate the credibility we have for a new prediction;
3) When making a new prediction, the nonconformity measurement α^* for the test sample x^* is computed based on the previous equation. Then, the P-value of the prediction, which indicates the credibility of the prediction is calculated by investigating the fraction of samples in the training distribution with larger nonconformity measurement:

$$p^{*,y} = \frac{\left\{i = 1,\ldots,n \middle| \alpha_i^{*,y} \le \alpha_i^n\right\}}{n} \tag{7}$$

Here $p^{*,y}$ is the P-value of the assumed label y for the new sample x^*. It should be noted that there are two P-values associated with the two labels (0: no ICU required, 1: need ICU) respectively.

4) The credibility of the prediction can be computed as the larger P-value [37], which reflects how well the most likely label conforms to the distribution of the training data nonconformity measurement. If the credibility is low, the credibility we have for the prediction is low, which may provide us with a flexible tool to tune the policies in medical resource management.

In this study, to show the validity of the reliability quantification with conformal prediction and whether more reliable predictions mean a higher likelihood of being correct, we used the conformal prediction with shrunken centroids (CPSC) algorithm and tested it on the HSS data set as an example (considering the superior performance on this dataset and the lowest computational time of SC) and performed two types of reliability analysis:

On the one hand, we partitioned the predictions into two categories based on the prediction credibility: unreliable predictions if the credibility is below a threshold ϵ, and reliable predictions if the credibility is above a threshold. Furthermore, within the reliable predictions, there are positive predictions and negative predictions. To make it simple to understand, the output of the conformal predictor is described as a tri-light system: red (reliable positive predictions), yellow (unreliable predictions), and green (reliable negative predictions). Then, we investigated the variations in AUROC, AUPRC, F1-score, and the number of predictions within the red- and green-light predictions as the credibility threshold ϵ changes.

On the other hand, we investigated the credibility of the correct prediction, the false positive predictions, and the false negative predictions and tested whether the correct predictions were assigned with higher credibility values.

For both analyses, to show the model performance robustness, we bootstrapped the training data 30 times and reported the metrics within the 30 parallel experiments.

3.7. Ethics and registration

The protocol of this multi-center study was approved by the institutional review board of Jinling Hospital, Nanjing University School of Medicine (2020NZKY-005–02). The written informed consent was waived because this was a retrospective study that presented no more than minimal risk of harm to subjects and involved no such procedures.

4. Results

4.1. Patient cohort

We collected 3646 patients for analysis, including a training cohort (n = 1451), a validation set (n = 662), an external test set (n = 1263) based on the data collected from the Huoshenshan field hospital, and a specific test set (n = 544) based on Delta and Omicron variants Fig. 2. Prediction models were built for the prediction of ICU admission (adverse cases in training set/validation set/HSS test set/Variants test set, n = 39/21/262/11, respectively). This cohort had 1830 men (50.2 %) and 1816 women (50.8 %), with a median age of 53.7 years (IQR, 42–65 years). The median age among men was 53.2 years (IQR, 41–65 years) and the median age among women was 52 years (IQR, 44–65 years). No statistical difference in age was found between men and women in this cohort.

4.2. Model prediction performance

The performance of the prediction models on the validation set, Huoshenshan (HSS) test set and the Delta and Omicron variants test set is shown in Table 1 and Fig. 3. According to the results shown in Fig. 3 when the models were trained without bootstrapping, the three models generally performed well on the validation set and on the two test sets under data distribution drifts when the models were trained on the training data originated from the first wave of COVID outbreak. The flexible LGB model performs the best on the validation set (AUROC: 0.89, AUPRC: 0.42) and the Delta and Omicron variant dataset (AUROC: 0.85, AUPRC: 0.60), while the simpler SC model performs the best on the HSS dataset (AUROC: 0.77, AUPRC: 0.51).

Table 1
The performance of the ICU prediction models on three data sets over 30 bootstrapping experiments. Mean and 95 % confidence intervals are reported.

Data set	Model	Feature Selection	Classifier Hyperparameter	AUROC	AUPRC
Validation	SC	LASSO $C = 0.01$	$\Delta = 0.01$	0.803 [0.793, 0.813]	0.221 [0.192, 0.250]
Validation	LGB	SC $\Delta = 0.05$	$\alpha = 1$, $D = 9$, $N = 15$	0.861 [0.849, 0.873]	0.306 [0.278, 0.334]
Validation	ANN	SC $\Delta = 0.4$	layers = [50,30,20]	0.755 [0.713, 0.800]	0.156 [0.090, 0.256]
HSS	SC	LASSO $C = 0.01$	$\Delta = 0.01$	0.743 [0.734, 0.753]	0.462 [0.448, 0.477]
HSS	LGB	SC $\Delta = 0.05$	$\alpha = 1$, $D = 9$, $N = 15$	0.705 [0.696, 0.715]	0.438 [0.427, 0.449]
HSS	ANN	SC $\Delta = 0.4$	layers = [50,30,20]	0.731 [0.636, 0.788]	0.431 [0.340, 0.513]
Variants	SC	LASSO $C = 0.01$	$\Delta = 0.01$	0.739 [0.719, 0.760]	0.418 [0.388, 0.448]
Variants	LGB	SC $\Delta = 0.05$	$\alpha = 1$, $D = 9$, $N = 15$	0.787 [0.773, 0.801]	0.518 [0.496, 0.542]
Variants	ANN	SC $\Delta = 0.4$	layers = [50,30,20]	0.804 [0.684,0.869]	0.552 [0.360,0.666]

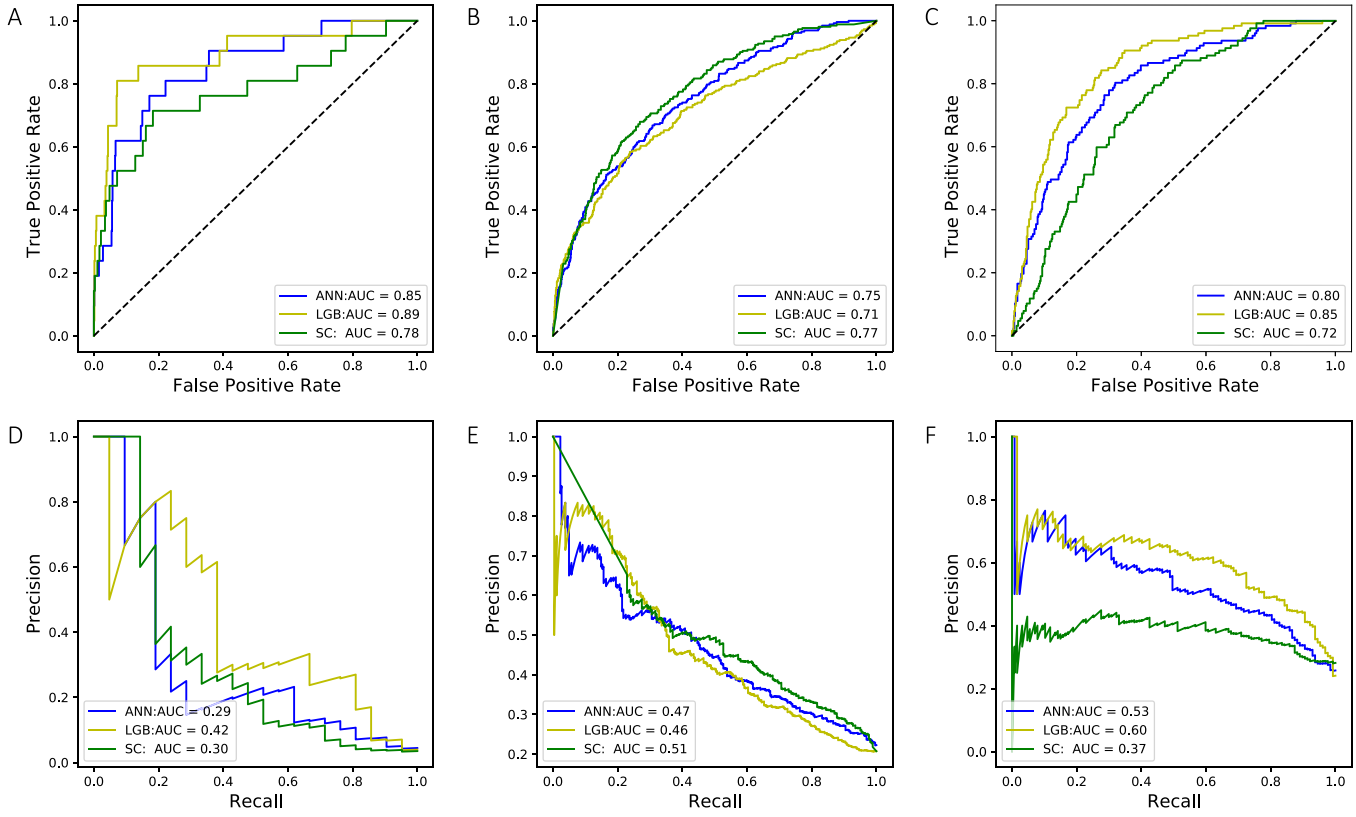

Fig. 3. The receiver operating characteristic curves (ROC) and precision-recall curves (PRC) of the models evaluated on the validation set (A, D), the Huoshenshan field hospital data set (B, E), and the Delta and Omicron variant data set (C, F).

4.3. Important feature analysis

To investigate the important features and their influence on the model decision, we leveraged the SC algorithm considering its high AUROC and AUPRC on the HSS dataset under data distribution drift, lowest computational time as well as its simplicity and interpretability based on the shrunken contrast introduced in Section 3D. After bootstrapping the training data 30 times and taking a mean value of the shrunken contrast (d'_+ for the positive predictions and d'_- for the negative predictions) over the 30 parallel experiments, firstly, we report the relative feature importance of the three types of data when viewed holistically: lab test data, clinical data and radiomic data in Fig. 4. The results indicate that, when these three types of features are compared, the average feature importance of the lab test data is higher than that of the other two types of features ($p < 0.01$, Wilcoxon rank-sum test). However, while the majority of the radiomics features have relatively lower shrunken contrast, there have been outliers among the radiomics features bearing high importance relatively higher than those of most of the lab test data.

Then, besides comparing the feature importance of the three major types, to specify the most important individual features across all features, the features with an absolute value of shrunken contrast above 0.4 ($d'_\pm > 0.4$) and the signs of the features associated with a positive prediction are reported in Table 2. A positive sign indicates a positive association with the outcome. The results show that the most important feature found by SC is the presence of dyspnea and the sign is positive, which indicates that patients with dyspnea are more likely to be predicted as needing ICU admission. besides the clinical symptom of dyspnea and the lab test value of Lactate dehydrogenase, the majority of the important features are based on radiomics which suggests that certain CT radiomic features are the main features that are important in the model decision-making process. Although the importance of CT radiomics is relatively less evident than the lab test features when viewed as a group, there are still many features important from CT radiomics in the model decision-making process.

4.4. Prediction credibility analysis

To assist in flexible medical resource management, besides giving the

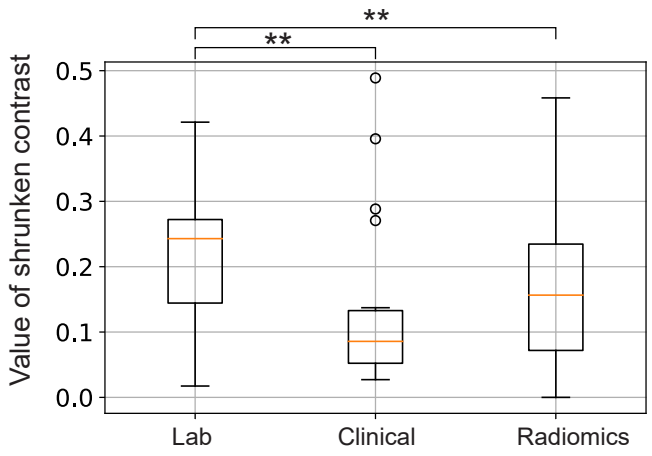

Fig. 4. The feature importance quantified by the absolute values of the shrunken contrasts for the positive class centroid averaged over 30 times of bootstrapping experiments. The lab test data contains 19 features; the clinical data contains 18 features; the radiomic data contains 9913 features. Statistical significance: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

predictions, the conformal prediction framework was applied to enable the users to understand the reliability of each specific prediction made by the algorithms. The results of the two types of reliability analyses introduced in Section 2 F are shown in Fig. 6. The results indicate that as the credibility threshold increases, the number of predictions decreases as the unreliable predictions with credibility below the threshold are filtered out. Meanwhile, the model performance metrics: AUROC, AUPRC, and F1-score increase accordingly when the more reliable predictions are investigated. The trend indicates that as the credibility requirement becomes stricter, the prediction performance is improved at the sacrifice of the number of predictions made by the algorithm: when the threshold is set higher, the AUROC, AUPRC, and F1-score are generally higher for those predictions confidently made by the model but as a sacrifice, the number of predictions that the model can confidently make decreased. To balance the model efficiency to making more predictions and the prediction reliability, an $\epsilon = 0.3$ can be a relatively balanced choice.

In addition, the mean credibility of the correct predictions, false positive predictions, and false negative predictions are reported in Fig. 5. According to the results, the mean credibility of the correct predictions is significantly higher than those of the false positive predictions and false negative predictions ($p < 0.001$, Wilcoxon rank-sum test). To sum up, the two types of reliability analyses show that credibility is positively related to prediction performance: the higher the prediction credibility is, the more likely the prediction is correct, which enables the users to flexibly control the prediction performance via prediction credibility Fig. 6.

5. Discussion and conclusion

Since the outbreak of COVID-19, the virus has continuously mutated. While most variants have shown reduced virulence compared to the initial strain, there have been sudden outbreaks of mutant strains with increased infectiousness (e.g., Omicron) or morbidity (e.g., Delta)[3,50,51]. The WHO classifies these strains as "variants of concern"[3]. The sudden emergence of these variants often overwhelms medical resources, such as general wards and intensive care units (ICUs)[41]. Our Tri-light Warning System was developed to predict ICU admissions. Results demonstrate that the system is reliable and effective under data distribution shifts, exhibiting strong performance with both the original strain (AUROC: 0.89, AUPRC: 0.42) and the Omicron and Delta variants (AUROC: 0.77–0.85, AUPRC: 0.51–0.60). Additionally, by employing a conformal prediction framework, we gained valuable insights into the reliability of predictions. This enables users to make informed decisions regarding ICU resource allocation based on the credibility of the predictions. Therefore, the Tri-light Warning System is an effective tool for stratifying COVID-19 patients under varying data conditions and can potentially enhance responses in future pandemics.

The performance evaluation of severity prediction models becomes particularly crucial when considering different mutational variants of COVID-19. Our findings indicate successful validation of the ICU prediction models, namely LGB, SC, and ANN, on distinct datasets: the Huoshenshan field hospital test set and the Delta and Omicron variant test set. Among these models, the LGB model demonstrated superior performance on the validation set (originating from the initial COVID outbreak), as well as on the Delta and Omicron variant dataset. Interestingly, in our ICU inpatients, the rate of ICU admissions among COVID-19-boosted individuals was statistically higher in the Omicron-infected group compared to the Delta-infected group, aligning with data highlighting the Omicron variant's immune-evasive properties [52–54]. Another possible reason for this is that the Omicron data came from a well-known and well-established general hospital in China that tended to admit patients with more pronounced symptoms, more comorbidities, or who had not been vaccinated during the Omicron wave, which might contribute to the increased ICU admission rate

Table 2
The most important features found by the shrunken centroids algorithms and their signs for making a positive prediction. The features were selected based on an absolute shrunken contrast above 0.4 averaged over 30 bootstrapping experiments. For the radiomic features, the last field indicates the statistics calculated over all lesions for a patient (e.g., 25: 25th percentile, 75: 75th percentile).

Feature Name	Feature Type	Absolute Shrunken Contrast	Sign
Dyspnea	Clinical	0.489	+
wavelet-LLL_firstorder_Minimum_25	Radiomics	0.458	−
wavelet-LLL_firstorder_Minimum_mean	Radiomics	0.455	−
wavelet-LLH_glrlm_RunEntropy_25	Radiomics	0.442	+
wavelet-LHH_glrlm_RunEntropy_median	Radiomics	0.440	+
wavelet-LHH_glrlm_RunEntropy_mean	Radiomics	0.436	+
wavelet-LHH_glrlm_RunEntropy_75	Radiomics	0.435	+
wavelet-LHH_glrlm_RunEntropy_25	Radiomics	0.434	+
wavelet-LLH_glrlm_RunEntropy_median	Radiomics	0.431	+
wavelet-LLL_firstorder_Minimum_median	Radiomics	0.423	−
Lactate dehydrogenase	Lab test	0.421	+
wavelet-LHH_glszm_ZoneEntropy_25	Radiomics	0.417	+
wavelet-LLH_glrlm_RunEntropy_mean	Radiomics	0.412	+
wavelet-LLH_glszm_ZoneEntropy_mean	Radiomics	0.410	+
lbp−3D-k_gldm_LargeDependenceHighGrayLevelEmphasis_75	Radiomics	0.409	+
wavelet-LHH_glszm_ZoneEntropy_median	Radiomics	0.408	+
wavelet-LLL_glszm_HighGrayLevelZoneEmphasis_mean	Radiomics	0.404	+
wavelet-LHH_glszm_ZoneEntropy_mean	Radiomics	0.404	+

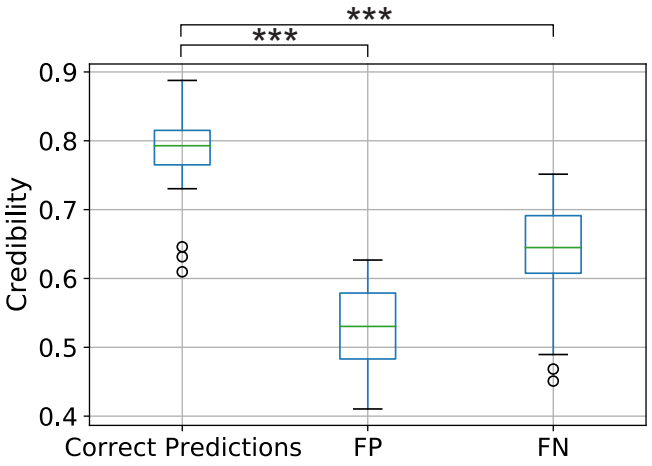

Fig. 5. The mean credibility of different types of predictions. Correct predictions, false negative predictions (FN), and false positive predictions (FP) are reported. The mean credibility is calculated with data in the same prediction types in each bootstrapping experiment (30 times of bootstrapping experiments in total). Statistical significance: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

among Omicron patients in our dataset. In contrast, the simpler and more interpretable shrunken centroids (SC) model performed best on the Huoshenshan dataset, which involved data drift due to the field hospital setting. This comparison between models is noteworthy. The more flexible LGB and ANN models outperformed the simpler SC models on the original training cohort and the variant dataset, representing data collected from hospitals with permanent resources (the validation dataset and the Delta and Omicron dataset). These models showcase the potential to generalize to newer variants. However, it is important to consider the data distribution drift caused by different types of hospitals. The flexible models may overfit the training data and exhibit higher variance when applied to the dataset collected from the Huoshenshan field hospital. Conversely, the SC model consistently performed well, indicating its ability to generalize across different types of hospitals without succumbing to data distribution overfitting, as observed in datasets collected from hospitals with permanent resources.

To identify discriminating features for prediction, we found that the presence of dyspnea, the lab test lactate dehydrogenase, three first-order radiomic features, and 13 higher-order (glrlm, glszm, gldm-based) radiomic features are important in the model decision making. Lactate dehydrogenase (LDH), a liver biochemistry marker related to liver impairment, has been shown to be associated with poor prognosis for patients with COVID-19 [55,56]. The presence of dyspnea has been widely recognized as an indicator of the severity of COVID-19 and the potential admission to ICU. It has also been deemed by the Centers for Disease Control and Prevention as a severe symptom of COVID-19 as well as for any COVID vaccines. In our previous study [7] where an LGB model was used to analyze the important features in the prediction of ICU requirements, mechanical ventilation requirements, and whether a patient would die from COVID-19 within 28 days, the majority of the important features are clinical features and lab test features including change of LDH and the presence of dyspnea [7]. However, in this study, the majority of important features found by the SC are radiomic features. We hypothesize that more radiomic features being important may be the main reason that enables the model to perform better than LGB under

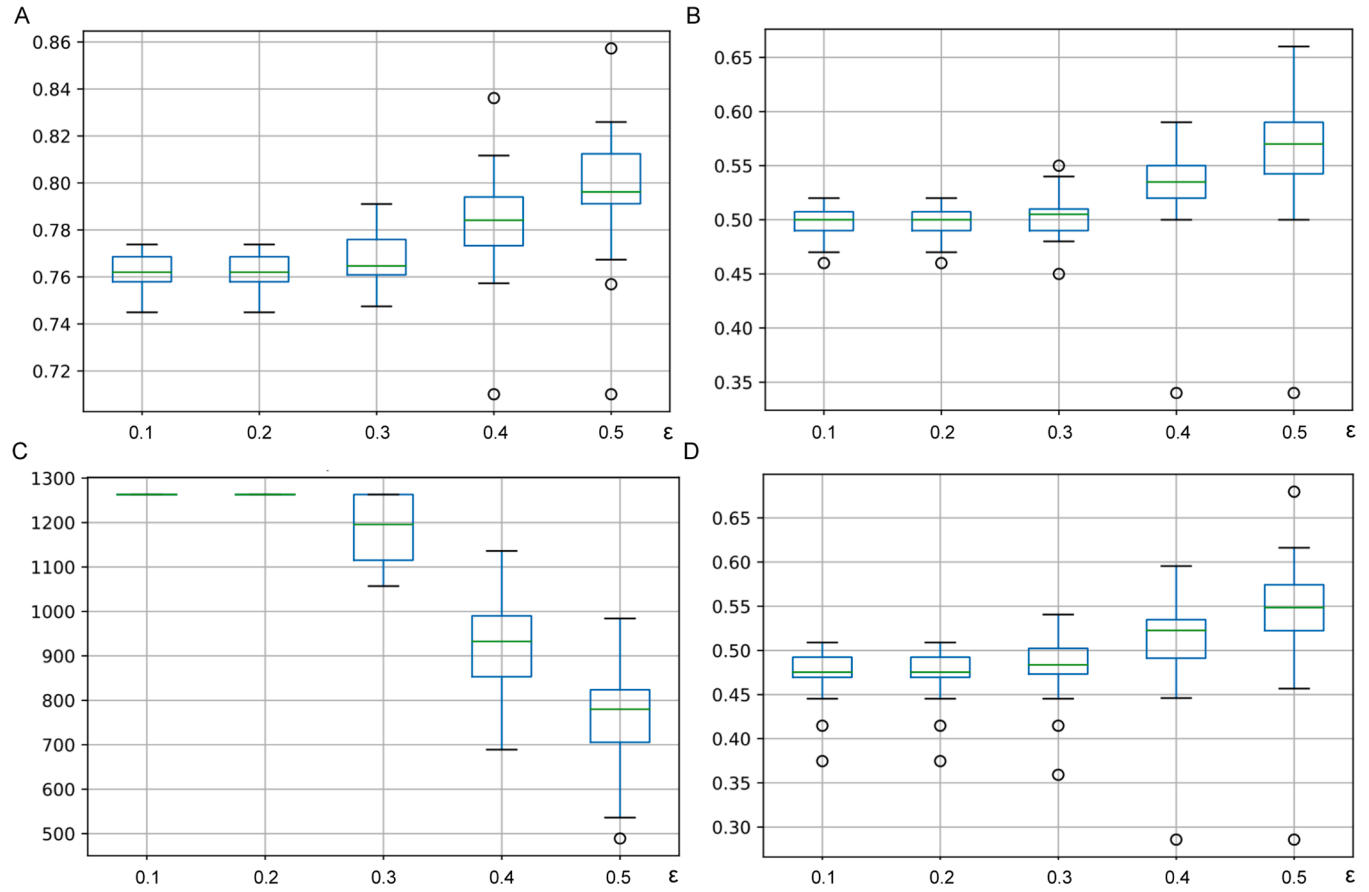

Fig. 6. The model performance under the varying threshold of credibility. The AUROC (A), AUPRC (B), number of predictions (C) made, and F1-score (D) after the unreliable predictions are filtered out based on the varying credibility threshold ϵ. It should be noted that 30 results from the bootstrapping experiments were reported.

data distribution drifts. Considering the difference between Huoshenshan field hospitals and other hospitals in the handling of lab tests (e.g. within 3d after admission vs within 24 h after admission) and the different clinical symptoms brought about by different COVID variants, the CT radiomic features are potentially more homogenous across different hospitals or for different COVID variants. Therefore, we suggest that CT radiomic features may be more robust under data distribution drifts.

Conformal prediction enables the model to output not only the predicted labels but also the credibility of the predictions [12,15]. We must clarify that the prediction reliability is not equal to the predicted probability. While predicted probability can be an intuitive indicator of a model's confidence in a particular prediction [48], it relies heavily on model assumptions and does not give uncertainty information explicitly. By contrast, the credibility given by the conformal prediction better reflects prediction uncertainty because it relies on the statistical distribution of the nonconformity measurement in the training observations with the i.i.d. assumption that is less strict than the assumptions of most machine learning models. Changing the threshold on the prediction credibility is also relatively independent of changing the threshold on the predicted probability. The thresholding on credibility can be used to quantify the confidence a user has in a prediction and whether or not to believe in the predictions made by the predictor, while the thresholding on predicted probability reflects what types of predictions are made by the model for a specific patient. To sum up, prediction credibility and prediction probability can be regarded as two key types of prediction information. Based on the quantified credibility, a clinical center can flexibly adapt its health policies based on both the model predictions and the prediction credibility.

In addition to the prediction credibility offered by the conformal prediction framework, we have introduced a tri-light warning system called "Red-Yellow-Green." This system enhances resource allocation by categorizing patients into different labels. Red-label patients are prioritized for ICU resources[57], while yellow-label patients require closer monitoring to assess their medical resource needs. Green-label patients are deemed stable with a low likelihood of disease progression. Moreover, the tri-light system, built upon prediction credibility, can adapt to varying virus variants.

It is important to note that patients predicted as "yellow" are not necessarily safer than those labeled as "red". The tri-light system explicitly indicates to the healthcare team that "the current medical information is insufficient to make a reliable judgment for the prediction labeled yellow". Thus, additional medical monitoring should be maintained. It is worth emphasizing that the dynamic nature of the tri-light system, utilizing the conformal prediction framework, enables the adaptive setting of health policies based on various scenarios, considering specific resource availability and emerging pandemic variants. Therefore, we propose this flexible model with double thresholds that can be adaptively determined by medical institutions in real-world settings.

While this study presents a predictive model for ICU resource allocation in COVID-19 patients, it is important to acknowledge several limitations. Firstly, the current model is designed solely to predict the need for ICU care and does not specifically predict the requirement for specific medical resources such as mechanical ventilation or extracorporeal membrane oxygenation (ECMO) in severe patients. Despite our efforts to predict ICU mortality based on updated data upon ICU admission (with an AUROC of 0.77 and AUPRC of 0.06 on the HSS test set), the model's performance was suboptimal due to limited training data. In the future, as more ICU cases are collected from various medical institutions, it will be worthwhile to try to develop more detailed patient outcome prediction models and further optimize medical resource allocation after ICU admission. Secondly, this study does not account for the effect of treatment. For instance, patients receiving appropriate treatments such as oxygen therapy, antiviral medications, and antibiotics may experience quicker recovery and improved outcomes, potentially avoiding the need for ICU admission despite being predicted to require it. The dataset used in this study includes variations in treatment protocols, which may arise from differences in location or changes in treatment guidelines over time. These factors may contribute to differences within the dataset. In the future, the treatment information (treatment type, healthcare provider information, etc.) could be encoded into the feature vector as another modality, which may potentially improve the patient stratification efficacy.

Furthermore, this study simulates the real-world scenario of presenting new COVID-related medical data to medical institutes using previously trained models. We assess the generalizability of the ICU prediction models under potential data drifts without supervised learning on the test sets. While testing and validating the models trained on previous data, we raise awareness of model generalizability issues across different medical facilities and COVID variants. It's important to note that these models are not intended for final and universal use by any institute. In the future, when new labeled data becomes available and model prediction performance becomes a priority, re-training the models, employing transfer learning and domain adaptation, instead of simply changing the credibility threshold, can address data drifts and ensure an up-to-date model before implementation.

Moreover, despite the utilization of extensive sample sizes and the inclusion of distinct prognosis information across varied data distributions, it is important to note that our study exclusively gathered data from hospitals located in China. This geographical restriction may potentially curtail the applicability of our models to other regions due to divergent protocols employed across different countries. In the future, the model effectiveness should further be validated on the dataset collected from multiple countries.

Furthermore, in our study, we acknowledge the assumption that the training distribution of nonconformity measurement may not adequately represent the ground-truth distribution. Ideally, for effective conformal prediction, two distinct sets should be present: 1) a proper training set to train the nonconformity measurement computation algorithm/model, and 2) a calibration set to evaluate the p-values of the test nonconformity measurements. However, due to the limited data available in our study, we combined these two sets for testing purposes, which may result in overfitting. In future research, as more data becomes available, particularly in terms of positive cases, it is recommended to partition the training data into separate proper training data and calibration data sets. This partitioning will help mitigate the potential overfitting issue associated with conformal prediction.

To conclude, we have introduced an end-to-end AI system that utilizes conformal predictions to efficiently identify high-risk COVID-19 patients. This system provides prediction probabilities and assesses the credibility of each prediction, classifying patients into high-risk, low-risk, and uncertain-risk categories. Our approach is adaptable to both the original strain of COVID-19 and its emerging variants. In the future, this system could play a crucial role in addressing future pandemic challenges, optimizing medical resource allocation, and enhancing disease management efficiency. Additionally, the validity of the model should be tested on multiple datasets across institutes and across different countries.

Code availability

The codes that support the findings of this study are available here: https://github.com/LeoLee7/COVID_trilight.

Funding statement

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (grant No. 82202150 to Xiao Li).

Ethical statement

The institutional review board approved this retrospective investigation and was in line with the Health Insurance Portability and Accountability Act. The need for informed patient consent was waived.

CRediT authorship contribution statement

Zhen Zhou: Writing – review & editing, Supervision, Conceptualization. **Mu Zhou:** Writing – review & editing, Supervision, Conceptualization. **Zijian Xing:** Writing – review & editing, Software. **Xianghao Zhan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chuanjun Xu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Olivier Gevaert:** Writing – review & editing, Supervision, Conceptualization. **Qinmei Xu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Guangming Lu:** Writing – review & editing, Supervision, Data curation, Conceptualization. **Li Liu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xiao Li:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Changsheng Zhou:** Supervision. **Danyang Ren:** Writing – review & editing, Conceptualization. **Longjiang Zhang:** Writing – review & editing, Supervision, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Xiao Li reports financial support was provided by National Natural Science Foundation of China. NA reports a relationship with NA that includes:. NA has patent NA pending to NA. NA If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available on request from the corresponding author (G.M.L.).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejro.2024.100603.

References

[1] W. H. Organization, Covid-19 epidemiological update: Global overview and sars-cov-2 variants, ⟨https://www.who.int/publications/m/item/covid-19-epidemiological-update-edition-167⟩, 2024, accessed: 2024-05-17.

[2] S. Jaumdally, M. Tomasicchio, A. Pooran, A. Esmail, A. Kotze, S. Meier, L. Wilson, S. Oelofse, C. van der Merwe, A. Roomaney, et al., Frequency, kinetics and determinants of viable sars-cov-2 in bioaerosols from ambulatory covid-19 patients infected with the beta, delta or omicron variants, Nat. Commun. 15 (1) (2024) 2003.

[3] D. Mallick, A. Singh, E.Y.-K. Ng, V. Arora, Classifying chest x-rays for covid-19 through transfer learning: a systematic review, Multimed. Tools Appl. (2024) 1–60.

[4] S. Turbow, T. Walker, S. Culler, M.K. Ali, Care fragmentation and readmission mortality and length of stay before and during the covid-19 pandemic: data from the national readmissions database, 2018–2020, BMC Health Serv. Res. 24 (1) (2024) 622.

[5] S. Imtiaz, E.M. Batubara, M.H. Abuelgasim, M.M. Alabad, L.M. Alyousef, N. H. Alqahtani, A.Y. Sabbagh, F.A. Alharbi, A.S. Ibrahim, Long-term outcome of pulmonary involvement in patients with coronavirus disease 2019: The role of high-resolution computed tomography and functional status–a prospective single-center observational study, Ann. Thorac. Med. 19 (2) (2024) 147–154.

[6] S.P. Koyyada, T.P. Singh, A systematic survey of automatic detection of lung diseases from chest x-ray images: Covid-19, pneumonia, and tuberculosis, SN Comput. Sci. 5 (2) (2024) 229.

[7] Q. Xu, X. Zhan, Z. Zhou, Y. Li, P. Xie, S. Zhang, X. Li, Y. Yu, C. Zhou, L. Zhang, et al., Ai-based analysis of ct images for rapid triage of covid-19 patients, NPJ Digit. Med. 4 (1) (2021) 1–11.

[8] H. Chen, M.-J. Chen, L.-Q. Ling, J.-R. Yang, H.-X. Huang, J.-J. Zhou, N. Yang, M.-J. Zhang, A prediction model for clinical outcomes of covid-19 hospitalized patients: construction and accuracy assessment, Clin. Lab. 70 (3) (2024).

[9] H. Ergenc, Z. Ergenc, S. Araç, I.H. Tor, E. Alkılınç, Predictors of disease severity, clinical course, and therapeutic outcome in covid-19 patients: our experience with 1700 patients, Author Prepr. (2024).

[10] I.S. Farahat, A. Sharafeldeen, M. Ghazal, N.S. Alghamdi, A. Mahmoud, J. Connelly, E. van Bogaert, H. Zia, T. Tahtouh, W. Aladrousy, et al., An ai-based novel system for predicting respiratory support in covid-19 patients through ct imaging analysis, Sci. Rep. 14 (1) (2024) 851.

[11] P.G. Asteris, S. Kokoris, E. Gavriilaki, M.Z. Tsoukalas, P. Houpas, M. Paneta, A. Koutzas, T. Argyropoulos, N.F. Alkayem, D.J. Armaghani, et al., Early prediction of covid-19 outcome using artificial intelligence techniques and only five laboratory indices, Clin. Immunol. 246 (2023) 109218.

[12] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer-Verlag, New York, 2005, https://doi.org/10.1007/b106715.

[13] X. Zhan, Z. Wang, M. Yang, Z. Luo, Y. Wang, G. Li, An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction, Measurement 158 (2020) 107588.

[14] H. Wieslander, P.J. Harrison, G. Skogberg, S. Jackson, M. Fridén, J. Karlsson, O. Spjuth, C. Wählby, Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images, IEEE J. Biomed. Health Inform. 25 (2) (2020) 371–380.

[15] X. Zhan, X. Guan, R. Wu, Z. Wang, Y. Wang, Z. Luo, and G. Li, Online conformal prediction for classifying different types of herbal medicines with electronic nose, 2018.

[16] X. Zhan, F. Wang, O. Gevaert, Reliably filter drug-induced liver injury literature with natural language processing and conformal prediction, IEEE J. Biomed. Health Inform. 26 (10) (2022) 5033–5041.

[17] X. Zhan, Q. Xu, Y. Zheng, G. Lu, and O. Gevaert, Reliability-based cleaning of noisy training labels with inductive conformal prediction in multi-modal biomedical data mining, arXiv preprint arXiv:2309.07332, 2023.

[18] A. Husnain, H.K. Hussain, H.M. Shahroz, M. Ali, A. Gill, S. Rasool, Exploring ai and machine learning applications in tackling covid-19 challenges, Rev. Esp. De. Doc. Cient. 18 (02) (2024) 19–40.

[19] M. Tamala, M.M. Rahmanb, M. Alhasimc, M.A. Mulhimd, and M. Derichee, Artificial intelligence (ai) based prediction of mortality, for covid-19 patients, arXiv preprint arXiv:2403.19355, 2024.

[20] E.D. Tenda, J. Henrina, A. Setiadharma, D.J. Aristy, P.Z. Romadhon, H. F. Thahadian, B.A. Mahdi, I.M. Adhikara, E. Marfiani, S.D. Suryantoro, et al., Derivation and validation of novel integrated inpatient mortality prediction score for covid-19 (impact) using clinical, laboratory, and ai—processed radiological parameter upon admission: a multicentre study, Sci. Rep. 14 (1) (2024) 2149.

[21] S. Yildirim, O. Sunecli, C. Kirakli, Mortality prediction with machine learning in covid-19 patients in intensive care units: a retrospective and prospective longitudinal study, J. Crit. Intensive Care-Vol. 15 (1) (2024) 31.

[22] W.D. de Holanda, L.C. e Silva, Á.A.d.C.C. Sobrinho, Machine learning models for predicting hospitalization and mortality risks of covid-19 patients, Expert Syst. Appl. 240 (2024) 122670.

[23] T. Hata, T. Goto, S. Yamanaka, T. Matsumoto, O. Yamamura, H. Hayashi, Prognostic value of initial serum sodium level in predicting disease severity in patients with covid-19: a multicenter retrospective study, J. Infect. Chemother. 30 (3) (2024) 181–187.

[24] A. Ramón, A. Bas, S. Herrero, P. Blasco, M. Suárez, J. Mateo, Personalized assessment of mortality risk and hospital stay duration in hospitalized patients with covid-19 treated with remdesivir: a machine learning approach, J. Clin. Med. 13 (7) (2024) 1837.

[25] S. Wei, W. Xiaqin, L. Liwei, Z. Fasu, P. Ying, T. Pingping, Y. Furong, Analysis of risk factors for death in the coronavirus disease 2019 (covid-19) population: Data analysis from a large general hospital in anhui, china, Cureus 16 (5) (2024).

[26] F.M. Ince, O. AlkanBilik, H. Ince, Evaluating mortality predictors in covid-19 intensive care unit patients: Insights into age, procalcitonin, neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, and ferritin lactate index, Diagnostics 14 (7) (2024) 684.

[27] A. Imzil, O. Mansoury, A. Oulahbib, A. Latifa, S. Hind, Comparative study of the severity of covid-19 infection between female and male patients, Niger. Med. J. 65 (1) (2024) 56–66.

[28] M.G. Choi, Y.W. Joo, M.-H. Kim, S. Park, Y.-Y. Shin, E.M. Chun, Prognostic factors for predicting post-covid-19 condition in patients with covid-19 in an outpatient setting, J. Korean Med. Sci. 39 (2) (2024).

[29] Y. Shi, X. Yu, H. Zhao, H. Wang, R. Zhao, J. Sheng, Host susceptibility to severe covid-19 and establishment of a host risk score: findings of 487 cases outside wuhan, Crit. care 24 (1) (2020) 1–4.

[30] M. Lucijanic, I. Krecak, I. Busic, A. Atic, J. Stojic, A. Sabljic, E. Soric, P. Veic, S. Marevic, L. Derek, et al., Estimated plasma volume status in covid-19 patients

and its relation to comorbidities and clinical outcomes, J. Thromb. Thrombolysis 57 (1) (2024) 50–57.
[31] N.J. Kyala, I. Mboya, E. Shao, F. Sakita, K.G. Kilonzo, L. Shirima, A. Sadiq, E. Mkwizu, N. Chamba, A. Marandu, et al., Neutrophil-to-lymphocyte ratio as a prognostic indicator in covid-19: evidence from a northern tanzanian cohort, 2024–02, medRxiv (2024), 2024–02.
[32] R. Najjar, M.Z. Hossain, K.A. Ahmed, M.R. Hasan, Exploring machine learning strategies in covid-19 prognostic modelling: A systematic analysis of diagnosis, classification and outcome prediction, 2024–03, medRxiv (2024), 2024–03.
[33] B. Ji, L. Kong, J. Wang, C. Liu, K. Yuan, L. Zhu, H. Liang, Predicting the prognosis of patients with mild covid-19 by chest ct based on machine learning, Chin. J. Acad. Radiol. (2024) 1–7.
[34] X. Fang, F. Shi, F. Liu, Y. Wei, J. Li, J. Wu, T. Wang, J. Lu, C. Shao, Y. Bian, Tracheal computed tomography radiomics model for prediction of the omicron variant of severe acute respiratory syndrome coronavirus 2, Die Radiol. (2024) 1–10.
[35] Z. Albataineh, F. Aldrweesh, M.A. Alzubaidi, Covid-19 ct-images diagnosis and severity assessment using machine learning algorithm, Clust. Comput. 27 (1) (2024) 547–562.
[36] M. Yin, C. Xu, J. Zhu, Y. Xue, Y. Zhou, Y. He, J. Lin, L. Liu, J. Gao1, X. Liu, D. Shen, and C. Fu, Automated machine learning for the identification of asymptomatic covid-19 carriers based on chest ct images, ⟨https://link.springer.com/article/10.1186/s12880-024-01211-w⟩, 2024, accessed: 2024-02-27.
[37] X. Zhan, F. Wang, O. Gevaert, Reliably filter drug-induced liver injury literature with natural language processing and conformal prediction, IEEE J. Biomed. Health Inform. 26 (10) (2022) 5033–5041.
[38] K. Drukker, W. Chen, J. Gichoya, N. Gruszauskas, J. Kalpathy-Cramer, S. Koyejo, K. Myers, R.C. Sá, B. Sahiner, H. Whitney, et al., Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment, J. Med. Imaging 10 (6) (2023) 061104.
[39] L. Wynants, B. Van Calster, G.S. Collins, R.D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D.L. Dahly, J.A. Damen, T.P. Debray, et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, bmj (2020) 369.
[40] R.F. Wolff, K.G. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, P. Group†, Probast: a tool to assess the risk of bias and applicability of prediction model studies, Ann. Intern. Med. 170 (1) (2019) 51–58.
[41] L. Rasmy, M. Nigo, B.S. Kannadath, Z. Xie, B. Mao, K. Patel, Y. Zhou, W. Zhang, A. Ross, H. Xu, et al., Recurrent neural network models (covrnn) for predicting outcomes of patients with covid-19 on admission to hospital: model development and validation using electronic health record data, Lancet Digit. Health (2022).
[42] D. Yamasoba, K. Uriu, A. Plianchaisuk, Y. Kosugi, L. Pan, J. Zahradnik, J. Ito, K. Sato, Virological characteristics of the sars-cov-2 omicron xbb. 1.16 variant, Lancet Infect. Dis. (2023).
[43] T. Liang, et al., Handbook of covid-19 prevention and treatment, First Affil. Hosp., Zhejiang Univ. Sch. Med. Compil. Clin. Exp. 68 (2020).
[44] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proc. Natl. Acad. Sci. 99 (10) (2002) 6567–6572.
[45] L. Liu, X. Zhan, X. Yang, X. Guan, R. Wu, Z. Wang, Z. Luo, Y. Wang, G. Li, Cpsc: Conformal prediction with shrunken centroids for efficient prediction reliability quantification and data augmentation, a case in alternative herbal medicine classification with electronic nose, IEEE Trans. Instrum. Meas. 71 (2022) 1–11.
[46] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).
[47] R. Collobert and S. Bengio, Links between perceptrons, mlps and svms, in Proceedings of the twenty-first international conference on Machine learning, 2004, 23.
[48] A.N. Angelopoulos and S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv preprint arXiv:2107.07511, 2021.
[49] H. Wang, X. Zhan, L. Liu, A. Ullah, H. Li, H. Gao, Y. Wang, R. Hu, G. Li, Unsupervised cross-user adaptation in taste sensation recognition based on surface electromyography, IEEE Trans. Instrum. Meas. 71 (2022) 1–11.
[50] T.P. Peacock, C.M. Sheppard, J.C. Brown, N. Goonawardane, J. Zhou, M. Whiteley, T.I. de Silva, W.S. Barclay, P.V. Consortium, et al., The sars-cov-2 variants associated with infections in india, b. 1.617, show enhanced spike cleavage by furin, BioRxiv (2021).
[51] N. Kumar, S. Quadri, A.I. AlAwadhi, M. AlQahtani, Covid-19 recovery patterns across alpha (b. 1.1. 7) and delta (b. 1.617. 2) variants of sars-cov-2, Front. Immunol. 379 (2022).
[52] A. Piralla, F. Mojoli, L. Pellegrinelli, F. Ceriotti, A. Valzano, G. Grasselli, M. R. Gismondo, V. Micheli, A. Castelli, C. Farina, et al., Impact of sars-cov-2 omicron and delta variants in patients requiring intensive care unit (icu) admission for covid-19, northern italy, december 2021 to january 2022, Respir. Med. Res. 83 (2023) 100990.
[53] W. Dejnirattisai, J. Huo, D. Zhou, J. Zahradník, P. Supasa, C. Liu, H.M. Duyvesteyn, H.M. Ginn, A.J. Mentzer, A. Tuekprakhon, et al., Sars-cov-2 omicron-b. 1.1. 529 leads to widespread escape from neutralizing antibody responses, Cell 185 (3) (2022) 467–484.
[54] F. Lyngse, C. Kirkeby, M. Denwood, L. Christiansen, K. Mølbak, C. Møller, et al., Household transmission of sars-cov-2 omicron variant of concern subvariants ba. 1 and ba. 2 in denmark, Nat. Commun. 13 (1) (2022) 5760.
[55] S. Genc, A. Taghizadehghalehjoughi, M.E. Naldan, O. Gülcü, C. Caglayan, M. Spanakis, T.K. Nikolouzakis, A. Alegakis, A.O. Docea, A.I. Drocas, et al., Evaluation of various blood biomarkers associated with the outcomes of patients with covid-19 treated in intensive care units, Exp. Ther. Med. 27 (2) (2024) 1–11.
[56] W. Liang, H. Liang, L. Ou, B. Chen, A. Chen, C. Li, Y. Li, W. Guan, L. Sang, J. Lu, et al., Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19, JAMA Intern. Med. 180 (8) (2020) 1081–1089.
[57] R. Chakraborty, N. Achour, Setting up a just and fair icu triage process during a pandemic: a systematic review (MDPI), Healthcare 12 (2) (2024) 146.