



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method

Zahra Asghari Varzaneh^a, Azam Orooji^b, Leila Erfannia^c, Mostafa Shanbehzadeh^{d,*}

^a Department of Computer Science, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran

^b Department of Advanced Technologies, School of Medicine, North Khorasan University of Medical Science (NKUMS), North Khorasan, Iran

^c Department of Health Information Management, School of Management and Medical Informatics, Health Human Resources Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

^d Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran

ARTICLE INFO

Keywords:

COVID-19
Coronavirus
Artificial intelligent
Machine learning
Data mining
Intubation
Mechanical ventilator

ABSTRACT

Background: Predicting severe respiratory failure due to COVID-19 can help triage patients to higher levels of care, resource allocation and decrease morbidity and mortality. The need for this research derives from the increasing demand for innovative technologies to overcome complex data analysis and decision-making tasks in critical care units. Hence the aim of our paper is to present a new algorithm for selecting the best features from the dataset and developing Machine Learning (ML) based models to predict the intubation risk of hospitalized COVID-19 patients.

Methods: In this retrospective single-center study, the data of 1225 COVID-19 patients from February 9, 2020, to July 20, 2021, were analyzed by several ML algorithms which included, Decision Tree (DT), Support Vector Machine (SVM), Multilayer perceptron (MLP), and K-Nearest Neighbors (K-NN). First, the most important predictors were identified using the Horse herd Optimization Algorithm (HOA). Then, by comparing the ML algorithms' performance using some evaluation criteria, the best performing one was identified.

Results: Predictive models were trained using 12 validated features. Also, it found that proposed DT-based predictive model enables a reasonable level of accuracy (=93%) in predicting the risk of intubation among hospitalized COVID-19 patients.

Conclusions: The experimental results demonstrate the effectiveness of the proposed meta-heuristic feature selection technique in combining with DT model in predicting intubation risk for hospitalized patients with COVID-19. The proposed model have the potential to inform frontline clinicians with quantitative and non-invasive tool to assess illness severity and to identify high risk patients.

1. Background

The Coronavirus disease 2019 (COVID-19) also known as severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) has affected millions of people worldwide [1,2]. Approximately 15%–20% of symptomatic patient's onrush to serious complications such as severe pneumonia, Acute Respiratory Distress Syndrome (ARDS), cytokine storm syndrome, and Multi-system Organ Dysfunction (MOF) requiring Intensive Care Unit (ICU) hospitalization [3,4]. Many hospital systems face extreme challenges with the extraordinary number of critical cases, causing many ICU departments to reach or surpass capacity [5,6]. In response to this serious infection, the design and implementation of

predictive models will be essential to the optimal use of limited ICU resources and support for clinical decisions [7,8]. Physicians have reported problems in predicting the progression of COVID-19 in hospitalized patients, along with problems in diagnosing patients who are prone to rapid deterioration. This requirement is more accentuated, especially with regard to the unpredictability of the disease behavior and courses [9].

The COVID-19 patients who deteriorate and need critical support with their breathing, require mechanical ventilator [10]. Therefore there is an immediate demand for classifying cases to use respiratory intubation services [11]. Within a short span of the COVID-19 pandemic, many researchers have extensively interested in the introduction of new

* Corresponding author. Department of Health Information Technology, Ilam University of Medical Sciences, Ilam, Iran.

E-mail addresses: sh.40par@yahoo.com (Z.A. Varzaneh), leila.erfannia.2020@gmail.com (A. Orooji), Leila.erfannia@gmail.com (L. Erfannia), Mostafa.shanbehzadeh@gmail.com (M. Shanbehzadeh).

<https://doi.org/10.1016/j.imu.2021.100825>

Received 10 November 2021; Received in revised form 13 December 2021; Accepted 14 December 2021

Available online 28 December 2021

2352-9148/© 2021 The Authors.

Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

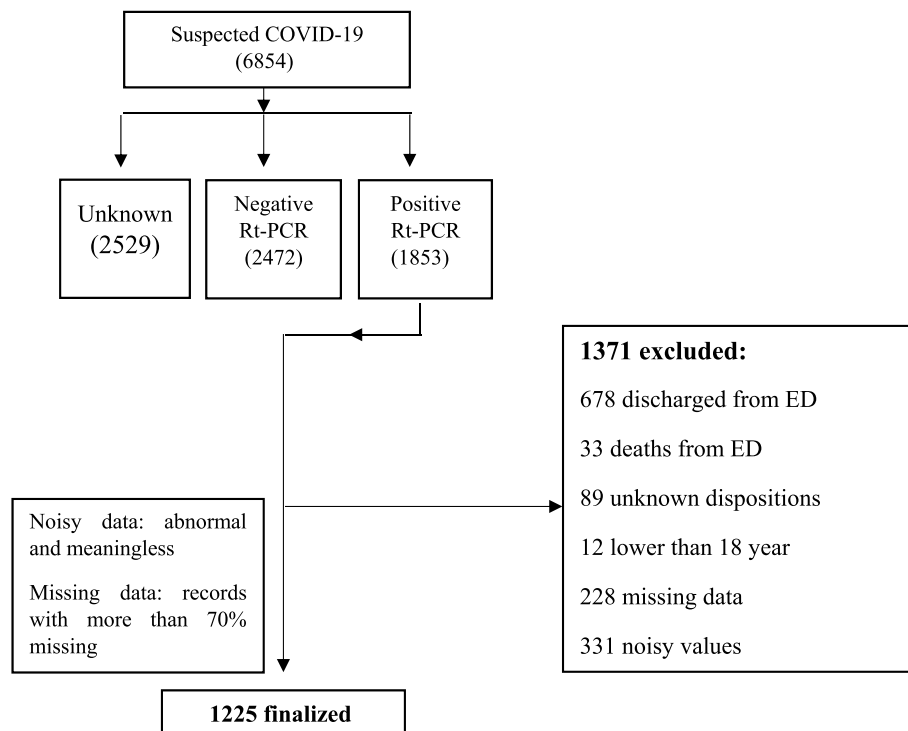


Fig. 1. Flow chart describing patient selection.

and non-invasive digital technologies such as Artificial Intelligence (AI) that can be effective in accurate and timely detection of patient at risk for clinical deterioration and severe hypoxia (low SPO2) [12,13]. It is proven that these methods may facilitate the identification of high-risk patients and adopt the most effective supportive oxygen therapy [13–15]. Machine Learning (ML) as a subfield of AI [16] is an essential tool for clinicians to identify patients at high risk for severe disease and to prioritize their hospitalization and resource utilization. Therefore, it can help reduce patient mortality and reduce the burden of health care resources(7, 17). In the prior studies, a large number of ML-based models were developed for predicting the risk of COVID-19 severity

and patient illness deteriorating [18,19], ICU admission [17,19–22], mechanical intubation [23] and deaths [17,20,24–29].

In this paper, we retrospectively analyzed the data of COVID-19 hospitalized patients from Imam Khomeini hospital, Ilam, Iran. At first, multiple meta-heuristic feature selection methods are compared based on the K-NN classification algorithm to select the best intubation predictors in patients with COVID-19. Then we construct and compare several ML-based prediction models for predicting the COVID-19 patients' severity requiring respiratory intubation based on selected variables. More precisely, the study questions posed for the experiment are: what are the most relevant predictors for predicting the COVID-19

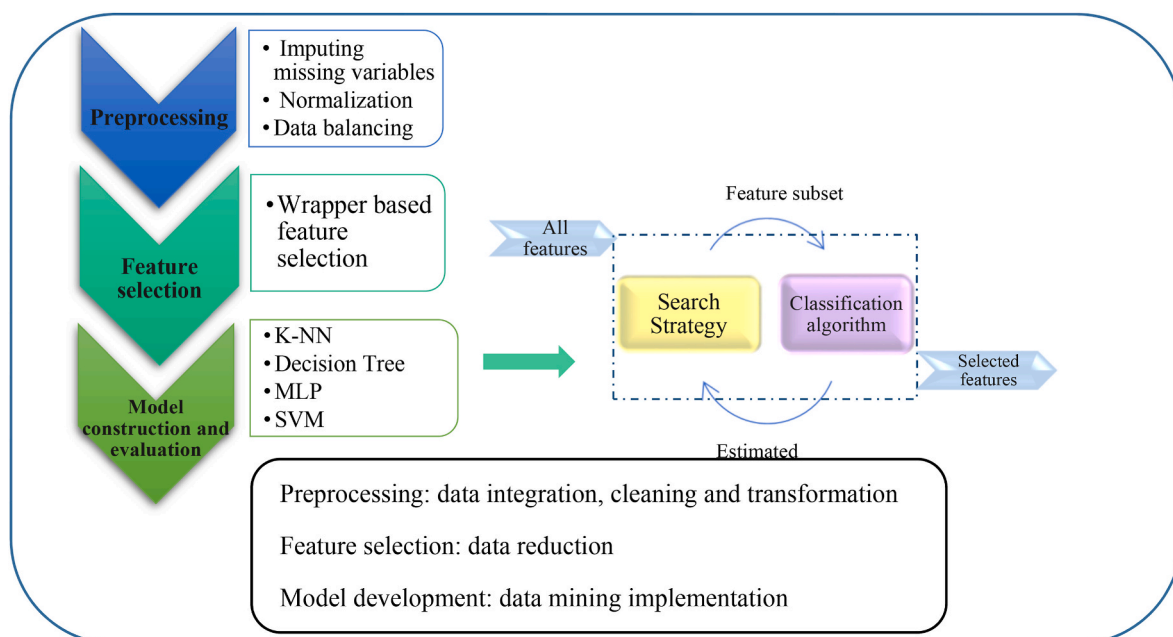


Fig. 2. An overview of proposed methodology.

intubation and 2- which prediction model presents better performance.

2. Material and methods

2.1. Data set description

In this study, a COVID-19 hospital-based registry data base from Imam Khomeini hospital, Ilam city, West of Iran, was retrospectively reviewed from February 9, 2020, to December 20, 2020. During this period, a total of 6854 suspected cases with COVID-19 had been referred to this center, of whom 1853 cases were introduced as positive COVID-19, 2472 as negative, and 2529 as unspecified. After applying exclusion criteria, for example negative RT-PCR COVID-19 test, unknown dispositions, discharged or death from emergency department, missing data >70%, noisy and abnormal values and out of defined time span, the number of 1225 case records were remained and their clinical data are recorded in the registry database based on 54 primary features (Fig. 1). These features are classified in six classes including patient's demographic (five features), clinical features (14 features), history of personal diseases (seven features), laboratory results (26 features), remedies (one feature) and an output variable (0: non-intubation and 1: intubation).

2.2. Study roadmap

The proposed method in this paper has three main phases. In the first phase, the raw data of patients with COVID-19 are preprocessed so that they can be used in the data mining process. In the second phase, which is our main goal in this study, effective clinical factors for predicting mortality in patients with COVID-19 are identified. In this way, with the help of metaheuristic optimization algorithms, the most important features that achieve higher predictive accuracy are extracted. After identifying the most important risk factors for intubation associated with COVID-19, ML algorithms are applied. Moreover, the performance of ML algorithms for prediction of intubation are evaluated and ranked based on statistical methods. The most important part of the proposed model is the feature selection, which works based on the rapper method. Details of the rapper-based model are shown in Fig. 2.

2.3. Dataset preprocessing

The use of COVID-19 raw data sets in the data mining process causes the efficiency of the algorithms to be low and the experimental results to be of poor quality. Useful information that can be extracted from the data directly affects our model's ability to learn. Therefore, it is very important to pre-process our data before inserting it into the model. Pre-processing is imperative to address irrelevant, redundant and unreliable data and to resolve inconsistencies [30]. In this regard, several pre-processing methods are used to prepare the data in order to use them for the data mining process.

Removing records with high missing rate improves the data mining accuracy and classification precision. It enhances learning efficiency, increasing predictive performance and reducing complexity of learning results [31,32]. In this paper, records with more than 70% of missing data were excluded from analysis. Finally, for the remaining records missing cells were imputed by mean and mode values substitution for continuous and discrete variables, respectively. The dataset used in this research are not balanced in terms of the number of records in each data class. This problem, disrupts the performance of ML algorithms. Hence, various techniques are introduced to deal with unbalanced datasets. In order to balance the data, we use the Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE produces synthetic samples of each minority class based on its nearest neighbors to increase the performance of the generalizer classifier on minority classes [33,34].

To manage noisy data, the normal range of each variable is defined using the opinion of two infectious disease specialist, two virologist and

hematologist. Then, we specified all the values that were outside the defined range and completed them by referring to the patient records or responsible doctor. Finally, normalization is used to reduce the variety of baseline between variables [35]. In this paper, MinMaxNormalization technique is used to scale dataset in (0, 1) interval as follow [33].

$$x = \frac{x - X_{min}}{X_{max} - X_{min}}$$

Where, X_{min} and X_{max} are maximum and minimum values of each column, respectively. This step is performed on KNIME analytics Platform.

2.4. Feature selection

Feature selection methods are one of the most important issues in ML and statistical pattern recognition [36,37] and divided into three main categories including filtering, wrapper and hybrid methods [38–40]. Feature selection in a high-dimensional data set is one of the most important steps in ML, which eliminates redundancy and irrelevant features in the dataset. So far, various classical methods have been proposed for the problem of feature selection, but with the magnification of real-world problems and the importance of the speed of access to the answer, the quality of their answer is not appropriate. The use of intelligent optimization methods has provided an effective help in solving these problems. Therefore, it can be said that one of the most effective and constructive problem solving algorithms is the selection of features and their dependencies, the use of meta-heuristic optimization methods and evolutionary algorithms [41,42]. Metaheuristic algorithms are a type of wrapper-based feature selection model in which ML algorithms are used to select optimal features. In other words, the criterion for selecting a feature is the efficiency of the classification algorithm.

In this paper, several well-known meta-heuristic algorithms including Horse herd Optimization Algorithm (HOA), Particle Swarm Optimization (PSO) [41], Genetic Algorithm (GA) [42], Grey Wolf Optimization (GWO) [42], Differential Evolution (DE) [43] and Harris Hawks Optimization (HHO) [44] have been used for feature selection. Feature selection can be considered as solution vectors whose length is equal to the total number of features and the values of its components are "0" or "1". A value of "1" indicates that the feature is selected and "0" is the non-selection of the feature. A "d" dimensional search space is limited to only these values in each dimension. Given that there are many modes of these solution vectors and as the number of features increase, the total number of solution modes increases exponentially, so the feature selection problem is an NP-Hard problem, and meta-heuristic algorithms are suitable for solving it. Besides the K-Nearest Neighbors (KNN) algorithm is used to calculate the fitness value of each of these algorithms in selecting the least number of features that have higher classification accuracy. The reason for using this algorithm is its simplicity of implementation and high classification accuracy. KNN can compete with the most accurate classification models and it is one of the most widely used machine learning algorithms, which is the best option for classification for most real-world problems. This algorithm is the best choice for problems where accuracy is important. Also, based on feature selection studies, the KNN algorithm is the best algorithm that can be used to calculate the fitness value [45–47].

The performance of the fitness value of meta-heuristic algorithms to find the features that have the highest classification accuracy with the least number of selected features is calculated by the following formula (Equation (1)).

$$\text{Fitness} = \alpha \gamma_R(D) + \beta \frac{|R|}{|C|} \quad (1)$$

Where $\gamma_R(D)$ is the classification error rate of KNN classifier. R is the number of selected features and C is the total number of features in the dataset, α and β are two parameters corresponding to the importance of classification quality and subset length. In this research, we set $\alpha = 0.99$

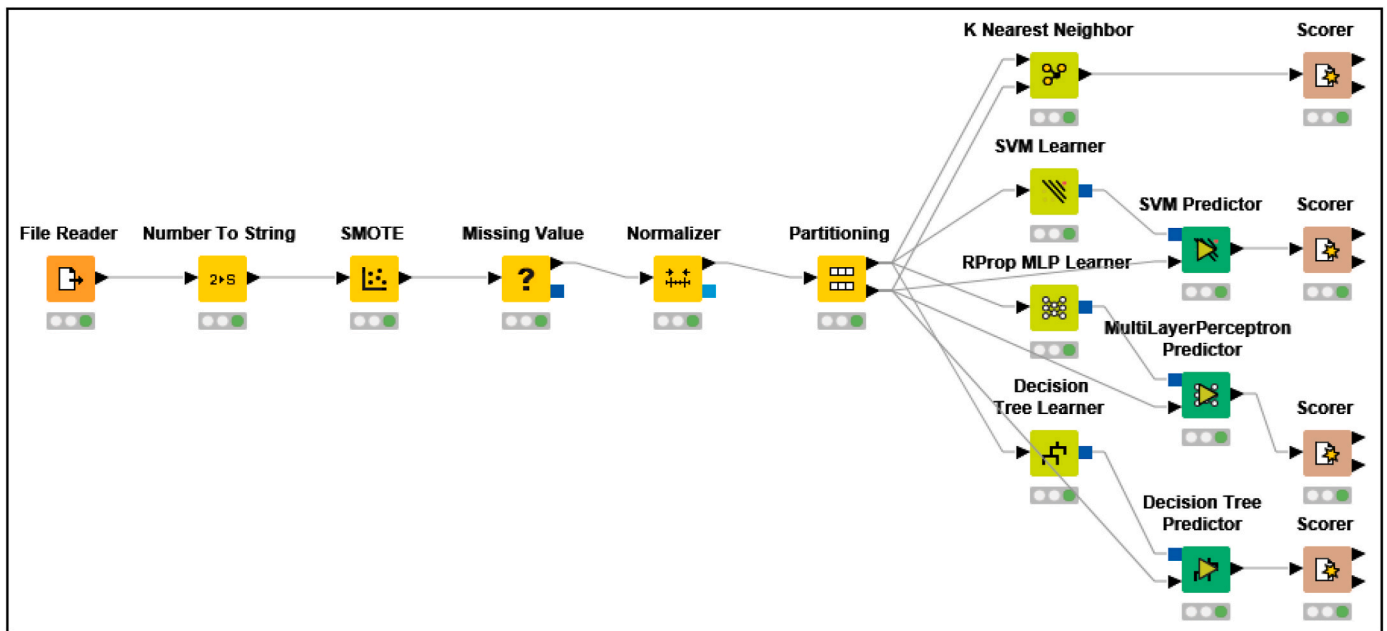


Fig. 3. KNIME workflow of classification methods.

Table 1
Definition of performance measures.

Performance Measures	Definitions
Precision	$TP / (TP + FP)$
Specificity/true negative rate (TNR)	$TN / (TN + FP)$
Sensitivity/true positive rate (TPR) or Recall	$TP / (TP + FN)$
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
F-measure	$(2 \times Precision \times Recall) / (Precision + Recall)$

*True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

and $\beta = 0.01$.

In this phase, all experiments were performed using MATLAB 2019 software. To evaluate the performance of meta-heuristic algorithms in identifying the most effective factors, three performance evaluation metrics are calculated: mean fitness value, classification accuracy using KNN, and the number of selected features.

2.5. Prediction model construction

As we can see in Fig. 3, the last step of the proposed method is to classify the data by the ML algorithms including K-NN, Multi-Layer Perceptron(MLP), Support Vector Machine(SVM) and Decision Tree (DT). In this step, ML algorithms are applied on dataset before and after running feature selection step. This step is performed on KNIME analytics Platform. The performance of classification algorithms was measured in terms of accuracy, precision, recall, specificity and F-

Table 2
Performance of ML algorithms before preprocessing and after preprocessing.

ML algorithm	Accuracy		Precision		Recall		Specificity		F-Measure		F.a.r p-value = 0.00544		F.a.r p-value = 0.00041	
	b.p	a.p	b.p	a.p	b.p	a.p	b.p	a.p	b.p	a.p	b.p	a.p	b.p	a.p
Decision tree	0.861	0.858	0.564	0.941	0.534	0.791	0.902	0.961	0.547	0.863	7.600			4.200
SVM	0.457	0.821	0.287	0.743	0.412	0.792	0.375	0.921	0.336	0.767	31.800			6.600
MLP	0.856	0.896	0.511	0.905	0.566	0.839	0.893	0.947	0.535	0.858	8.400			4.200
KNN	0.826	0.971	0.462	0.942	0.485	0.903	0.912	0.982	0.471	0.922	17.200			1.400

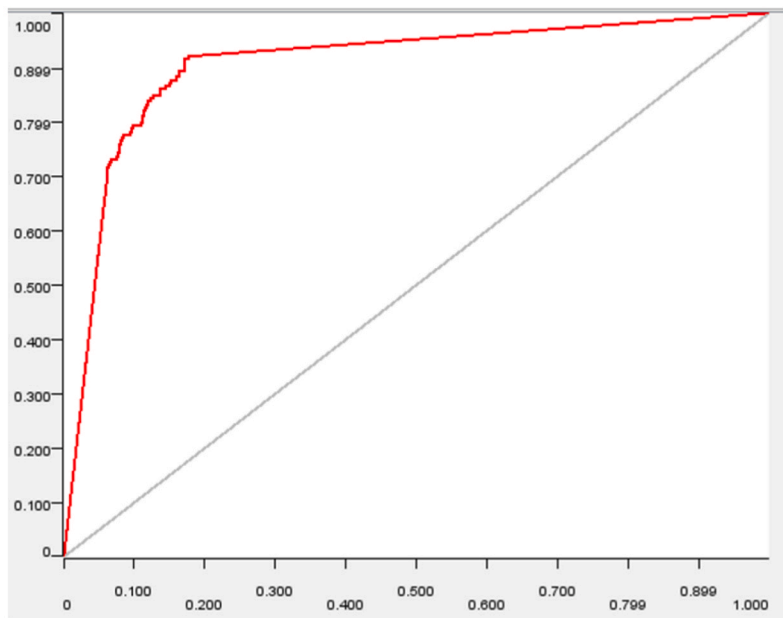
*b.p: before preprocessing, a.p: after preprocessing, F.a. r: Friedman aligned ranks.

measure. The calculation formula for each of these criteria is shown in Table 1. Also, in order to evaluate the performance of each classifier 10-fold Cross-Validation is used in which the data set was divided into 10 independent subsets and each subset was considered as test data and other data as training data. In addition, Friedman statistical technique was used to compare algorithms more precisely and select an algorithm with the highest efficiency. This test assigns a rank to each algorithm and the best algorithm has a lower rating. Hypothesis zero states that all algorithms are the same. While rejecting the null hypothesis shows that the compared algorithms are significantly different. In this paper, we set the significance level to $\alpha = 0.05$.

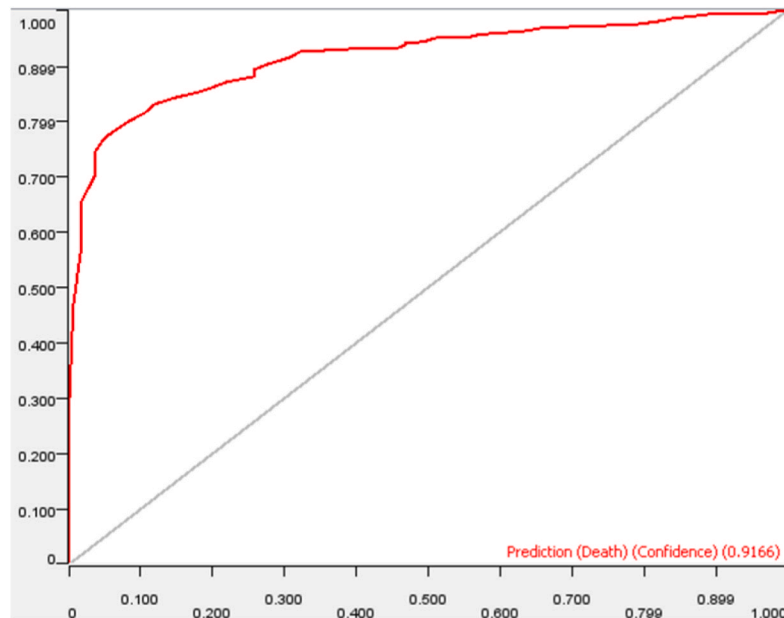
Fig. 3 shows the schematic workflow of this step which is performed on KNIME analytics Platform.

3. Results

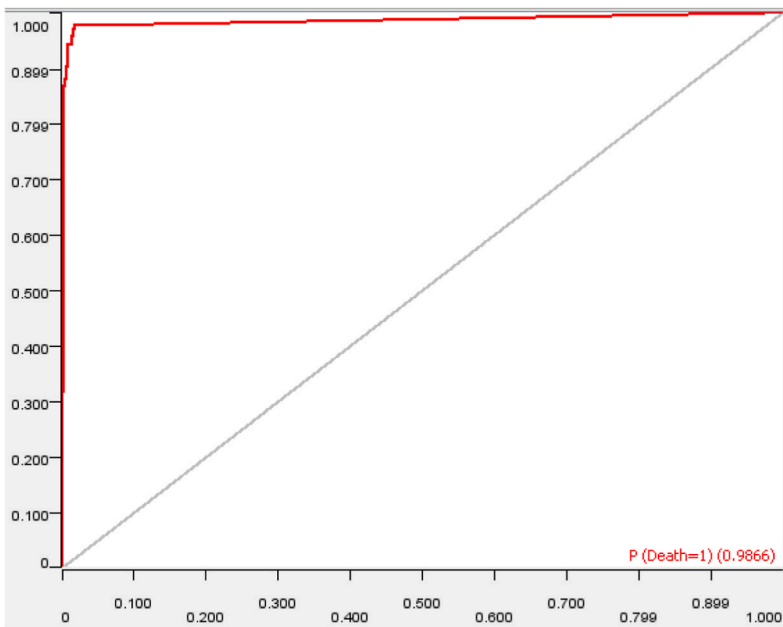
A total of 6854 suspected cases with COVID-19 had been referred to Imam Khomeini hospital which after applying the exclusion criteria and removing records with more than 70% missing data, records of 1225 positive RT-PCR patients remained. In order to balance the data, we use SMOTE method is used. Before data balancing, “intubation” class contained only 176 records (13%), while after balancing the dataset, number of records in this class raised to 748. Then, using Min-MaxNormalization technique all data normalized between 0 and 1. All ML algorithms were implemented on original and preprocessed dataset and the achieved results are shown in Table 2. Since, it is difficult to compare performance of ML algorithms based on five different evaluation metrics, the Friedman statistical test is used to compare and rank ML algorithms on the basis on these evaluation criteria. The results of



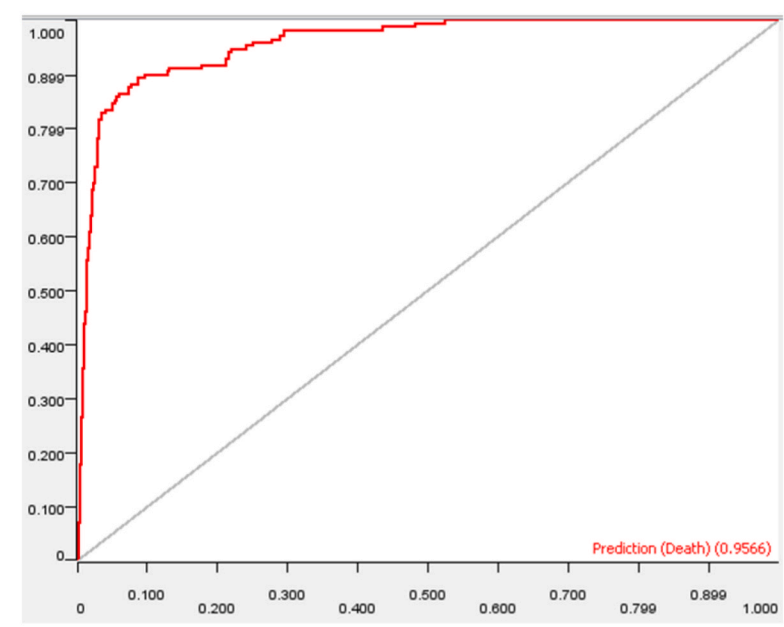
SVM



Decision tree



KNN



MLP

Fig. 4. ROC curve for ML algorithms.

Table 3
Comparison of algorithms in terms of different criteria in 20 runs.

Measure	Algorithms					
	GA	PSO	DE	GWO	HHO	HOA
Mean fitness value	0.101	0.095	0.096	0.098	0.101	0.083
Accuracy	0.891	0.903	0.904	0.900	0.892	0.924
No. selected features	17	21	20	24	19	12

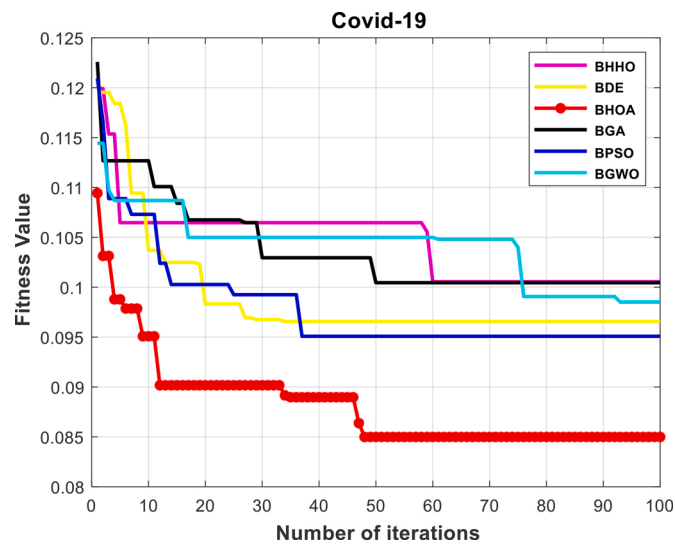


Fig. 5. Convergence comparison of meta-heuristic algorithms.

Friedman test with a significance level of 0.05 are also shown in Table 2. The achieved p-values indicate that there is a significant difference in performance of ML algorithms.

According to Table 2, the DT with rank 7.6 generally performed better than other algorithms for original data set. The results shown in Table 2 revealed that the performance of ML algorithms in prediction of the need to intubation has improved significantly after preprocessing. According to Table 2, KNN with mean rank of 1.4 had the best performance for preprocessed dataset. Additionally, ROC curves are plotted for all ML algorithms using preprocessed dataset. As shown in Fig. 4, KNN is the best model because the area under the ROC curve is close to 1.

Feature selection step is done by using binary version of well-known meta-heuristic algorithms including HOA, PSO, GA, GWO, DE and HHO. Due to the fact that meta-heuristic algorithms are random in nature and the solutions may be slightly different in each independent execution, so each algorithm is executed 20 times and finally, the average of the results are collected after 20 independent executions. Furthermore, in all algorithms, the population size and the maximum number of iterations are considered 50 and 100, respectively. The mean fitness value of each algorithm, the accuracy of the K-NN classifier based on the selected features, and the number of selected features are shown in Table 3.

Numerical results show that the HOA algorithm is superior to other algorithms in terms of all three criteria. We are looking for an algorithm that selects the least number of features and at the same time, can achieve higher classification accuracy. The HOA algorithm selects 12 of

the features as most effective risk factors for prediction of the need for intubation, including high age, high weight, dry cough, fever, dyspnea, loss of smell, cardiovascular diseases, hypertension, C-reactive protein, ALT/ASP, oxygen saturation (SPO₂), and leukocytosis. The convergence diagram of meta-heuristic algorithms is also shown in Fig. 5. Among meta-heuristic algorithms, the algorithm with the lowest fitness value has higher performance than other algorithms. The results show that the efficiency of HOA algorithm with fitness of 0.083 is higher than other algorithms in finding the least number of features with the highest classification accuracy.

After identifying the most important factors by the HOA algorithm, ML algorithms were used to predict the need of intubation. The obtained results from the implementation of these algorithms are presented in Table 4.

According to Table 4, DT with a rank of 1.2 is the best classification algorithm for predicting the need of patients with COVID-19 to intubation. The SVM algorithm with a mean rank of 4 is weaker than other algorithms. By comparing the best models in Tables 2 and 4, it is concluded that although the accuracy, precision and specificity are reduced but recall and F-measure are improved, after feature selection.

4. Discussion

Given the wide range of clinical manifestations of COVID 19, it is important to develop models for estimating the likelihood of intubation using ML techniques(7). In response to this life-threatening infection, the design and implementation of Clinical Decision Support Systems (CDSS) will be critical to the optimal use of hospital limited resources and support for clinical decisions. CDSS equipped ML can assist clinical decisions by informing caregivers and recommending interventions based on objective and generalizable empirical data [7,48,49].

In this article, we analyzed the data from a hospital registry database to develop and evaluate models capable of predicting the need for respiratory intubation in hospitalized COVID-19 patients according to baseline clinical features. First, the efficiency of six feature selection methods was compared to identify the best predictors. The results show that the efficiency of HOA algorithm is higher than others in finding the least number of features with the highest classification accuracy.

This study then adopted the most reliable and clinically relevant predictors related to intubation by using HOA method. Hence we identified 12 highly correlated variables with output class. Several studies selected some clinically important predictors for COVID-19 patient deterioration and progression of disease using the feature selection techniques. Selected features are used as input for the development of ML-based predictive models for the severity and intubation risk prediction in hospitalized COVID-19 patients. For example, Olmedo (2021) designed an intelligent CDSS based on some ML algorithms to predict future intubation among hospitalized patients with COVID-19. The four most relevant features for model prediction were Lactate Dehydrogenase (LDH) activity, CRP levels, neutrophil counts, and urea levels [50]. The most important variables in the Aljouie (2021) study for intubation prediction were age, BMI, LOS, oxygen saturation, D-dimer, and cardiovascular diseases [51]. Varun Arvind's(2021) results also showed that the laboratory variables of CRP, D-dimer, ALT, ASP, and leuko-/lymphocyte counts have a higher capability in predicting the intubation risk among hospitalized COVID-19 patients(23). Burdicka(2020) predicts the need for ventilation in COVID-19 by using some selected

Table 4
Performance evaluation results of classification algorithms.

ML algorithm	Accuracy	Precision	Recall	Specificity	F-Measure	Friedman aligned rank p-value = 0.00013
Decision tree	0.938	0.931	0.927	0.934	0.933	1.200
SVM	0.754	0.643	0.692	0.832	0.667	4.000
MLP	0.896	0.905	0.872	0.907	0.878	2.600
KNN	0.924	0.897	0.943	0.892	0.922	2.200

variables such as age, BMI, fever&chill, CRP, BUN, SPO2, lung lesion, and underline diseases [52]. In the present study, we used HOA meta-heuristic optimization technique to determine the weight of each predictor. Accordingly, the features of high age, high weight, dry cough, fever, dyspnea, loss of smell, cardiovascular diseases, hypertension, C-reactive protein, ALT/ASP, oxygen saturation (SPO2), and leukocytosis were the best predictive features.

ML can be applied to forecast future risk of mechanical intubation and may facilitate identification of high-risk patients to assist in clinical care [23]. Bolourani showed ML capable of accurate discrimination between COVID-19 patients at high risk versus low risk of requiring ventilation within 24 h [53]. The model predicted the need for mechanical ventilation using only routinely available labs and vital sign data.

It is proven that ML can be an effective tool in dealing with COVID-19 pandemic. Therefore we trained four well-known classification algorithms including DT, SVM, MLP, and RF classifiers according to the top 12 related parameters affecting the risk of intubation that derived from HOA feature selection. So far, several studies have been evaluating the application of ML techniques in predicting the COVID-19 poor outcomes. Olmedo (2021) designed an intelligent CDSS based on some ML algorithms to predict future intubation among hospitalized patients with COVID-19. Finally, the best performance was yielded by DT algorithm with AUC-ROC of 97%, accuracy of 94%, F-score of 77%, sensitivity of 93% and specificity of 95% [50]. Aljouie (2021) in their study, assessed the performance of four common ML algorithms such SVM, RF, Linear Regression (LR), and DT to model COVID-19 outcome prediction. Their results showed that the model developed using DT with 0.81 of AUC introduced as the best performing model [54]. Burdicka(2020) predict the need for respiratory ventilation in COVID-19 by using the DT classifier with sensitivity of 90%, specificity of 58%, and AUC of 86% [55].

The result of other studies confirm the better performance of DT than other similar algorithms in predicting the intubation of patients with COVID-19 (61,49,29,24,25). In this research, different data mining algorithms have been used to classify the data, which the DT algorithm has a higher efficiency than other algorithms. Accordingly in the current study, the results showed that the DT algorithm with precision of 0.931, recall of 0.927, specificity of 0.934, F-mesasure of 0.933 and accuracy of 0.938 has the best capability for early prediction of the risk of intubation in COVID-19 hospitalized patients. Based on the DT algorithm some clinical rules have been extracted, we have brought the two most important of them with the highest samples classified.

Rule 1: IF (oxygen saturation (SPO2) \leq 96% && pleural fluid = Yes, Activated partial thromboplastin time \leq 31) THEN the Intubation = True. This rule can be interpreted as overall among the 64 research samples who had more than 96% of oxygen saturation (SPO2), the 47 samples had the intubation process and the variable as the root node in the decision tree was considered as the most important factor for determining the endotracheal intubation risk among hospitalized COVID-19 patients.

Rule 2: IF (oxygen saturation (SPO2) $>$ 96% && dry cough = Yes && C-reactive protein $<$ 8.8 Mg/L && loss of smell = No) THEN the endotracheal intubation risk = negative. In this study, 221 samples had this rule template and among them, 187 samples have been classified correctly through this rule template as negative or low risk of endotracheal intubation. Generally, this rule with the most classified samples has been recognized as the most important decision rule in this research. Oxygen saturation as the most important node in our model was one the best predictor for intubation which in other studies also recognized as an important feature.

The results of the present study may help clinicians throughout correct, accurate and timely diagnosis of the disease progression and reduce the severe complications and the resulting mortalities. Despite the small amount of data fed into the models and the lack of some important clinical variables, the selected ML models, especially DT algorithm, performed well. On the other hand, this model application in

real clinical environments will assist physicians owing to its simplicity, user-friendliness and easy-to-use characteristics. Given the power of the current study in timely and accurate prediction of intubation risk, this study had some limitations that need to be addressed. First, this is a retrospective study that suffers from low data quantity (missing or duplicate cells) and non-optimal quality (imbalanced, noisy, and meaningless values). Second, we deal with a single-center dataset with limited sample size which undoubtedly confines the generalizability of the proposed model. Moreover, we used only four well known ML algorithms for prediction analyses based on some clinical features. Finally, the selected registry dataset lacks some important Para-clinical variables. In the future, the performance accuracy of our model and its generalizability will be enhanced if we test more ML techniques, at the larger, multicenter, and prospective dataset which is equipped with more qualitative and validated data.

5. Conclusion

The main idea behind this research is to evaluate several Meta-heuristic feature selection algorithms and ML models to predict future risk of intubation among hospitalized patients with COVID-19. The present study may assist medical specialist in choosing the optimal supportive oxygen therapy in the critically ill patient with respiratory failure through identification and prioritizing predictors and ML based predictions. Our developed prediction model has the potential to provide frontline physicians with an easy and fast tool to classify COVID-19 patients without having to wait for the results of additional tests. This predictive model also may be an advantage in better care delivery, lessen clinician workload, and diminish severe complication and death in the COVID-19 patients. In the future work, the proposed method is expected to be applied to other medical and healthcare domains such as early diagnosis and treatment of chronic disease. The meta-heuristic algorithms used in feature selection can also be improved.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This article is extracted from a research project supported by the Ilam University of Medical Sciences (IR.MEDILAM.REC.1399.294). We also thank the Research Deputy of the Ilam University of Medical Sciences for financially supporting this project.

References

- [1] Braam DH, Srinivasan S, Church L, Sheikh Z, Jephcott FL, Bukachi S. Lockdowns, lives and livelihoods: the impact of COVID-19 and public health responses to conflict affected populations - a remote qualitative study in Baidoa and Mogadishu, Somalia. *Conflict Health* 2021;15(1).
- [2] Heydariyan M, Mohammadtaghizadeh M, Shojaei M, Babazadeh M, Abbasian S, Amrovani M. The effect of COVID-19 derived cytokine storm on cancer cells progression: double-edged sword. *Mol Biol Rep* 2021;1–11.
- [3] Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* 2020;395(10229):1054–62.
- [4] Chow N, Fleming-Dutra K, Gierke R, Hall A, Hughes M, Pilishvili T. CDC COVID-19 Response Team. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019—United States, February 12–March 28, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69(13):382–6.
- [5] Kodama C, Kuniyoshi G, Abubakar A. Lessons learned during COVID-19: building critical care/ICU capacity for resource limited countries with complex emergencies in the World Health Organization Eastern Mediterranean Region. *J. global health* 2021;11:1–4.
- [6] Leclerc T, Donat N, Donat A, Pasquier P, Libert N, Schaeffer E, et al. Prioritisation of ICU treatments for critically ill patients in a COVID-19 pandemic with scarce resources. *Anaesthesia Critical Care and Pain Medicine* 2020;39(3):333–9.
- [7] Karthikeyan A, Garg A, Vinod P, Priyakumar UD. Machine learning based clinical decision support system for early COVID-19 mortality prediction. *Front Public Health* 2021;9.

- [8] Shanbezadeh M, Soltani T, Ahmadi M. Developing a clinical decision support model to evaluate the quality of asthma control level. *Middle East J Sci Res* 2013; 14(3):387–93.
- [9] Li X, Liao H, Wen Z. A consensus model to manage the non-cooperative behaviors of individuals in uncertain group decision making problems during the COVID-19 outbreak. *Appl Soft Comput* 2021:99.
- [10] Hawkins A, Stapleton S, Rodriguez G, Gonzalez RM, Baker WE. Emergency tracheal intubation in patients with COVID-19: a single-center, retrospective cohort study. *West J Emerg Med* 2021;22(3):678.
- [11] DBNet: a novel deep learning framework for mechanical ventilation prediction using electronic health records. In: Zhang K, Jiang X, Madadi M, Chen L, Savitz S, Shams S, editors. Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics; 2021.
- [12] Roshanaei G, Omid T, Faradmal J, Safari M, Poorolajal J. Determining affected factors on survival of kidney transplant in living donor patients using a random survival forest. *Koomesh* 2018;20(3):517–23.
- [13] Khan M, Mehran MT, Haq ZU, Ullah Z, Naqvi SR, Ihsan M, et al. Applications of artificial intelligence in COVID-19 pandemic: a comprehensive review. *Expert Syst Appl* 2021:185.
- [14] Asada K, Komatsu M, Shimoyama R, Takasawa K, Shinkai N, Sakai A, et al. Application of artificial intelligence in COVID-19 diagnosis and therapeutics. *J Personalized Med* 2021;11(9).
- [15] Rodríguez-Rodríguez I, Rodríguez JV, Shirvanizadeh N, Ortiz A, Pardo-Quiles DJ. Applications of artificial intelligence, machine learning, big data and the internet of things to the COVID-19 pandemic: a scientometric review using text mining. *Int J Environ Res Publ Health* 2021;18(16).
- [16] Kazemi A, Kazemi K, Sami A, Sharifian R. Identifying factors that affect patient survival after orthotopic liver transplant using machine-learning techniques. *Exp Clin Transplant* 2019;17:775–83.
- [17] Ryan L, Lam C, Mataraso S, Allen A, Green-Saxena A, Pellegrini E, et al. Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: a retrospective study. *Annals of Medicine and Surgery* 2020;59:207–16.
- [18] Assaf D, Gutman Ya, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Internal and emergency medicine* 2020;15(8):1435–43.
- [19] Agieb R. Machine learning models for the prediction the necessity of resorting to icu of covid-19 patients. *Int J Adv Trends Comput Sci Eng* 2020:6980–4.
- [20] Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. *PLoS One* 2020;15(7): e0236618.
- [21] Allenbach Y, Saadoun D, Maalouf G, Vieira M, Hellio A, Boddaert J, et al. Development of a multivariate prediction model of intensive care unit transfer or death: a French prospective cohort study of hospitalized COVID-19 patients. *PLoS One* 2020;15(10):e0240711.
- [22] Pan P, Li Y, Xiao Y, Han B, Su L, Su M, et al. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation [e23128 pp.]. *J Med Internet Res [Internet]* 2020;22(11). 2020/11//.
- [23] Arvind V, Kim JS, Cho BH, Geng E, Cho SK. Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. *J Crit Care* 2021;62:25–30.
- [24] Gao Y, Cai G-Y, Fang W, Li H-Y, Wang S-Y, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun* 2020;11(1):1–10.
- [25] Hernandez-Suarez DF, Ranka S, Kim Y, Latib A, Wiley J, Lopez-Candales A, et al. Machine-learning-based in-hospital mortality prediction for transcatheter mitral valve repair in the United States. *Cardiovasc Revascularization Med* 2021;22:22–8.
- [26] Parchure P, Joshi H, Dharmarajan K, Freeman R, Reich DL, Mazumdar M, et al. Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19. *BMJ Support Palliat Care* 2020.
- [27] Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR medical informatics* 2021;9(1): e24207.
- [28] Yadav AS, Li Y-c, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *The Lancet Digital Health* 2020;2(10):e516–25.
- [29] Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. *Nature machine intelligence* 2020;2(5):283–8.
- [30] García S, Luengo J, Herrera F. Data preprocessing in data mining. Springer; 2015.
- [31] Moyer CA, Compton SD, Kaselitz E, Muzik M. Pregnancy-related anxiety during COVID-19: a nationwide survey of 2740 pregnant women. *Arch Womens Ment Health* 2020;23(6):757–65.
- [32] Baalman SW, Lopes RR, Ramos LA, Neefs J, Driessen AH, van Boven WP, et al. Prediction of atrial fibrillation recurrence after thoracoscopic surgical ablation using machine learning techniques. *Diagnostics* 2021;11(10):1787.
- [33] Booth AL, Abels E, McCaffrey P. Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Mod Pathol* 2020:1–10.
- [34] Li L, Li B, Bai Y, Liu W, Wang H, Leung HC, et al. Abnormal resting state effective connectivity within the default mode network in major depressive disorder: a spectral dynamic causal modeling study. *Brain and behavior* 2017;7(7):e00732.
- [35] Hosseini GS, Nasrabadi AM. Effective connectivity of mental fatigue: dynamic causal modeling of EEG data. *Technol Health Care : official J. European Society Eng. Medicine* 2019;27(4):343–52.
- [36] Arnarson TO, Olason DT, Smári J, Sigurethsson JF. The Beck Depression Inventory Second Edition (BDI-II): psychometric properties in Icelandic student and patient populations. *Nord J Psychiatr* 2008;62(5):360–5.
- [37] Zhang Y, Xin Y, Li Q, Ma J, Li S, Lv X, et al. Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *Biomed Eng Online* 2017;16(1):125.
- [38] Dai L, Zhou H, Xu X, Zuo Z. Brain structural and functional changes in patients with major depressive disorder: a literature review. *PeerJ* 2019;7:e8170.
- [39] Rodrigues PA, Zaninotto AL, Neville IS, Hayashi CY, Brunoni AR, Teixeira MJ, et al. Transcranial magnetic stimulation for the treatment of anxiety disorder. *Neuropsychiatric Dis Treat* 2019;15:2743.
- [40] Li H, Chen Z, Gong Q, Jia Z. Voxel-wise meta-analysis of task-related brain activation abnormalities in major depressive disorder with suicide behavior. *Brain imaging and behavior* 2020;14(4):1298–308.
- [41] Molina-Ruiz RM, García-Saiz T, Looi JC, Virgili EV, Zamorano MR, de Anta Tejado L, et al. Neural mechanisms in eating behaviors: a pilot fMRI study of emotional processing. *Psychiatry investigation* 2020;17(3):225.
- [42] Park EY, Yi M, Kim HS, Kim H. A decision tree model for breast reconstruction of women with breast cancer: a mixed method approach. *Int J Environ Res Publ Health* 2021;18(7):3579.
- [43] Breast-Cancer detection using thermal images with marine-predators-algorithm selected features. In: Rajinikanth V, Kadry S, Taniar D, Damasevicius R, Rauf HT, editors. 2021 seventh international conference on bio signals, images, and instrumentation (ICBSII). IEEE; 2021.
- [44] Ghiasi MM, Zendeheboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. *Comput Biol Med* 2021;128:104089.
- [45] Shaban WM, Rabie AH, Saleh AI, Abo-Eloud M. A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. *Knowl Base Syst* 2020;205:106270.
- [46] EL-Hasnony IM, Elhoseny M, Tarek Z. A hybrid feature selection model based on butterfly optimization algorithm: COVID-19 as a case study. *Expert Syst* 2021: e12786.
- [47] Too J, Mirjalili S. A hyper learning binary dragonfly algorithm for feature selection: a COVID-19 case study. *Knowl Base Syst* 2021;212:106553.
- [48] Xia P, Gao K, Xie J, Sun W, Shi M, Li W, et al. Data mining-based analysis of Chinese medicinal herb formulae in chronic kidney disease treatment. *Evid Based Complement Alternat Med* 2020;2020:9719872.
- [49] Yazdani A, Safdari R, Ghazisaedi M, Beigy H, Sharifian R. Scalable architecture for telemonitoring chronic diseases in order to support the CDSSs in a common platform. *Acta Inf Med* 2018;26(3):195.
- [50] Domínguez-Olmedo JL, Gragera-Martínez Á, Mata J, Álvarez VP. Machine learning applied to clinical laboratory data in Spain for COVID-19 outcome prediction: model development and validation. *J Med Internet Res* 2021;23(4):e26211.
- [51] Aljouie AF, Almazroa A, Bokhari Y, Alawad M, Mahmoud E, Alawad E, et al. Early prediction of COVID-19 ventilation requirement and mortality from routinely collected baseline chest radiographs, laboratory, and clinical data with machine learning. *J Multidiscip Healthc* 2021;14:2017.
- [52] Burdick H, Lam C, Mataraso S, Siefkas A, Braden G, Dellinger RP, et al. Prediction of respiratory decompensation in Covid-19 patients using machine learning: the READY trial. *Comput Biol Med* 2020;124:103949.
- [53] Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. *J Med Internet Res* 2021;23(2):e24246.
- [54] Aljouie AF, Almazroa A, Bokhari Y, Alawad M, Mahmoud E, Alawad E, et al. Early prediction of COVID-19 ventilation requirement and mortality from routinely collected baseline chest radiographs, laboratory, and clinical data with machine learning. *J Multidiscip Healthc* 2021;14:2017–33.
- [55] Burdick H, Lam C, Mataraso S, Siefkas A, Braden G, Dellinger RP, et al. Prediction of respiratory decompensation in Covid-19 patients using machine learning: the READY trial. *Comput Biol Med* 2020;124.