BMJ Open

Machine learning techniques for mortality prediction in emergency departments: a systematic review

Amin Naemi ,[1] Thomas Schmidt ,[1] Marjan Mansourvar ,[1] Mohammad Naghavi-Behzad ,[2,3] Ali Ebrahimi ,[1] Uffe Kock Wiil [1]

To cite: Naemi A, Schmidt T, Mansourvar M, *et al*. Machine learning techniques for mortality prediction in emergency departments: a systematic review. *BMJ Open* 2021;**11**:e052663. doi:10.1136/bmjopen-2021-052663

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-052663).

Received 24 April 2021
Accepted 27 September 2021

[1]Maersk Mc-Kinney Moller Institute, Center for Health Informatics and Technology,University of Southern Denmark, Odense, Denmark
[2]Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[3]Department of Nuclear Medicine, Odense University Hospital, Odense, Denmark

Correspondence to
Amin Naemi;
amin@mmmi.sdu.dk

ABSTRACT

Objectives This systematic review aimed to assess the performance and clinical feasibility of machine learning (ML) algorithms in prediction of in-hospital mortality for medical patients using vital signs at emergency departments (EDs).

Design A systematic review was performed.

Setting The databases including Medline (PubMed), Scopus and Embase (Ovid) were searched between 2010 and 2021, to extract published articles in English, describing ML-based models utilising vital sign variables to predict in-hospital mortality for patients admitted at EDs. Critical appraisal and data extraction for systematic reviews of prediction modelling studies checklist was used for study planning and data extraction. The risk of bias for included papers was assessed using the prediction risk of bias assessment tool.

Participants Admitted patients to the ED.

Main outcome measure In-hospital mortality.

Results Fifteen articles were included in the final review. We found that eight models including logistic regression, decision tree, K-nearest neighbours, support vector machine, gradient boosting, random forest, artificial neural networks and deep neural networks have been applied in this domain. Most studies failed to report essential main analysis steps such as data preprocessing and handling missing values. Fourteen included studies had a high risk of bias in the statistical analysis part, which could lead to poor performance in practice. Although the main aim of all studies was developing a predictive model for mortality, nine articles did not provide a time horizon for the prediction.

Conclusion This review provided an updated overview of the state-of-the-art and revealed research gaps; based on these, we provide eight recommendations for future studies to make the use of ML more feasible in practice. By following these recommendations, we expect to see more robust ML models applied in the future to help clinicians identify patient deterioration earlier.

Strengths and limitations of this study

- ► To the best of our knowledge, this is the most recent systematic review on the topic of using machine learning (ML) algorithms to predict mortality at emergency departments.
- ► All included studies were evaluated regarding potential bias and the analysis was done following standard checklists.
- ► It was not possible to conduct meta-analysis, due to the heterogeneity of included studies.
- ► Only studies with vital signs among their predictors were included in the review.

INTRODUCTION

The improved life expectancy in Europe and most developed countries[1] has increased admissions to the emergency departments (EDs), especially for elderly people who have a higher risk of morbidity and mortality.[2 3] Also, having more intercity and intracity travels and organised emergency medical services to transfer these patients to ED brought more importance as well as patient load to EDs.[4 5] The average rate of mortality among in-patient admissions is around 2% in USA,[6] indicating the necessity of using clinical deterioration prediction models.[7] Moreover, a recent study has shown that some in-hospital deaths due to clinical deterioration can be avoided.[8]

Vital signs including heart rate (HR), pulse rate (PR), respiratory rate (RR), oxygen saturation (SpO_2), blood pressure (BP), body temperature (TP) and Glasgow coma scale (GCS) are essential parts of the patients' monitoring process. The data received from these parameters could be considered as the cheapest and the most important information, which could easily be collected during admission.[9] Although vital signs are among the first data that clinicians observed to have an initial assessment about the patients' condition and have been used in clinical practice for over a century, few studies have effectively quantified their performance in various clinical applications.[10] In recent years, various studies have emphasised that changes in vital signs happen several hours before a serious adverse event.[11–13] Nowadays, vital signs play a key role in identification of patients with

risk of deterioration in ED, however deterioration often occurs unnoticed or is not identified until it becomes too late to intervene.[9]

Machine learning (ML) addresses the question of how to build a computer model that automatically enhances performance through experiments, lying at the intersection of computer science, statistics and a variety of other disciplines concerned about automatic improvement over time, and inferring and decision-making under uncertainty.[14] Various ML algorithms have been developed to solve a wide variety of problems. The most commonly used ML algorithms are supervised learning algorithms focusing on relationships and dependencies between the input and output features.[15] Hence, we can identify the output values for unseen data based on the relationships learnt from the training data sets. Unsupervised learning is subset of ML that is used in pattern recognition and descriptive modelling without the use of output categorises or data labelling.[16]

Because of enormous advancements in the development of modern ML algorithms such as deep learning, the availability of large databases and increased computational power, ML has progressed rapidly over the last two decades from laboratory curiosity to practical technology in a wide range of commercial applications. Consequently, utilisation of ML techniques in the healthcare domain is increasing rapidly, and the list of tasks where ML has matched or outperformed physicians is growing. Clinical outcome prediction is one challenge receiving attention, and by improvement of ML techniques, the ML-based systems aim to outperform conventional clinical scoring systems.[14 17]

Various studies have been made in this area evaluating the prediction accuracy of in-hospital mortality, cardiac arrest and intensive care unit transfer. To reach this goal, various techniques such as logistic regression (LR), support vector machine (SVM), K-nearest neighbours (KNN), decision tree (DT), random forest (RF), Gaussian process (GP) and artificial neural networks (NN) have been applied, and the results confirmed the appropriateness of using ML in this area.[14] The capability of handling large data sets is the most considerable advantage of ML algorithms, which can result in more accurate prediction of clinical outcomes. Currently, we have various prediction models, using different techniques, in order to improve the triage process and to provide better identification of high-risk patients.[18 19] Although several studies based on different models were conducted, no single model has yet proved to be superior regarding clinical efficacy.[20–24]

Despite the use of vital signs in patient monitoring and clinical decision-making, the importance of specific types of vital signs, their correlation and frequency of registration to best prevent adverse events and in-hospital mortality is still unclear.[9] However, ML could be a good solution for these challenges. Considering the possible role of ML models in prediction of clinical deterioration of patients in EDs,[25] it is essential to analyse and merge the results of previous studies to have an updated overview of the state-of-the-art. This systematic review focused on the clinical efficacy and technical implications of ML-based models in prediction of mortality using vital signs predictors in patients admitted to EDs.

Table 1 Research questions

Q1	What ML techniques have been used to predict in-hospital mortality in EDs?
Q2	What are the common vital signs variables used in studies?
Q3	How researchers prepared data for ML techniques?
Q4	What are the approaches to solve the problem (eg, binary classification or time series prediction)?
Q5	What are the challenges and open issues in this domain?

EDs, emergency departments; ML, machine learning.

MATERIALS AND METHODS

This systematic review was conducted in the period from July 2020 to August 2021. The main questions addressed by this study are presented in table 1.

Search strategy and study selection

Three databases including Medline (PubMed), Scopus and Embase (Ovid) were targeted. Relevant articles were extracted using a broad range of relevant keywords. We stratified keywords into five groups namely, ML keywords, medical keywords, document type, publication year and language. The keywords in a group were paired using OR operators and all groups were paired using AND operator. Table 2 shows the search keywords that were applied to the titles, abstracts and full text in the three databases.

Table 2 Search keywords in different groups

Group 1—ML keywords	Artificial intelligence, machine learning, deep learning, learning algorithms, supervised machine learning, unsupervised machine learning
Group 2—Medical keywords	Clinical deterioration, mortality, in-hospital mortality, death, vital sign, emergency departments
Group 3—Document type	Journal
Group 4—Publication year	1 January 2010 to 1 August 2021
Group 5—Language	English
Final result	(Group 1) AND (Group 2) AND (Group 3) AND (Group 4) AND (Group 5)

ML, machine learning.

Naemi A, *et al*. *BMJ Open* 2021;**11**:e052663. doi:10.1136/bmjopen-2021-052663

Table 3 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Prediction of in-hospital mortality should be the main aim of the study	Studies which are conference articles, posters, abstracts, books or book chapters and review articles.
The study should be done at ED.	Studies which were done in paediatrics field.
The study should be done on adult patients.	Non-English studies.
ML algorithms should be used for the prediction task.	
Vital signs variables should be among the predictors used to build ML models.	

ED, emergency department; ML, machine learning.

Inclusion and exclusion criteria are listed in table 3. Our queries for three databases can be found in online supplemental file S1.

Data extraction

Three researchers (AN, MM and AE) screened the title and extracted abstracts, independently. The screening process was done using the Covidence tool. Then, two researchers (AN and TS) read the full texts, independently resolving the disagreement by supervision of the senior researcher (UKW). Spreadsheets for item extraction were prepared based on the critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS) checklist.[26] The extracted items include study design, publication date, source of data, study population, outcomes, time horizon, considered vital signs predictors, data preprocessing, model development, model performance, model validation and evaluation, and the interpretation.

Risk of bias assessment

Risk of bias (ROB) for each study was assessed by using the prediction risk of bias assessment tool (PROBAST) checklist[27] and reported based on an adapted form.[28] Extracted articles were classified into three different categories (low, high and unclear). A study is considered as having a high ROB if the study has at least one high ROB in four domains. ROB assessment of each study was checked by two of our researchers (AN and MN-B); disagreements were resolved in collaboration with a senior researcher (UKW).

Patient and public involvement

No patient involved.

RESULTS

The initial search resulted in 7466 records, which after removing duplicates, evaluation of eligibility for inclusion, full-text assessment, quality assessment and review of the references of the included papers, was narrowed down to fifteen studies. The study selection flowchart is shown in figure 1.

General information, as well as study design and population of included studies, are summarised in table 4. Twelve out of 15 articles were published after 2019, which indicates an increased interest recently among researchers to explore the capability of ML in prediction of clinical outcomes in EDs.

Studies design

Among extracted articles, 14 studies have used electronic data of the related hospitals, while 1 of the studies was based on registry-based public data.[29] One of the essential parts in utilising ML techniques in each application is gathering and preparation of the data set, which has a high impact on the performance of the final ML-based system. In general, ML techniques need an extensive data set to produce good results, and the performance of ML models depends largely on the number of training examples.[14] Based on Riley *et al*[30] 13 studies had enough number of patients to build acceptable models through their data sets, while 2 studies proposed a model based on only 100 and 445 patients.[31 32]

Vital signs predictors

We considered studies, which utilised vital signs among their predictors in building models. The vital signs set used in extracted articles included HR, PR, RR, SpO_2, BP, TP and GCS (table 4). BP and TP were the most widely used vital sign predictors among the included studies. SpO_2, RR and HR were common vital signs, which were used in 13, 13 and 11 articles, respectively, while GCS and PR were the least commonly used vital signs. Besides the vital signs, all studies used other clinical predictors such as demographic variables (eg, age, sex, race, marital status),[18 19 21 22 29 31–40] arrival mode (walk-in and ambulance),[18 19 29 32 36] blood tests (eg, albumin, creatinine, haemoglobin, potassium, sodium, white cell count, urea),[21 33 35–37 39] ECG signal,[31] chief complaints,[18 19 29 34 36 38 40] length of stay at hospital,[22] medications,[19 32 36–38] comorbidities,[36] diagnoses,[22 38] medical history,[18 36–38] and triage information.[19 21 36 38]

Data preparation

Real-world data sets often need to be transformed, cleaned or changed before use. These data sets may contain missing values, noise and even wrong entries or inconsistencies.[41] Missing values issue is one of the main challenges in clinical data mining[42] that often provide additional information about patients. Therefore, in almost all cases, raw data will need to be preprocessed before it can be used as the basis for ML modelling. This preprocessing may affect the final model, and the specific steps are important to clarify. However, 10 articles did not utilise proper techniques to impute missing values and excluded up to 32% of incomplete patients'

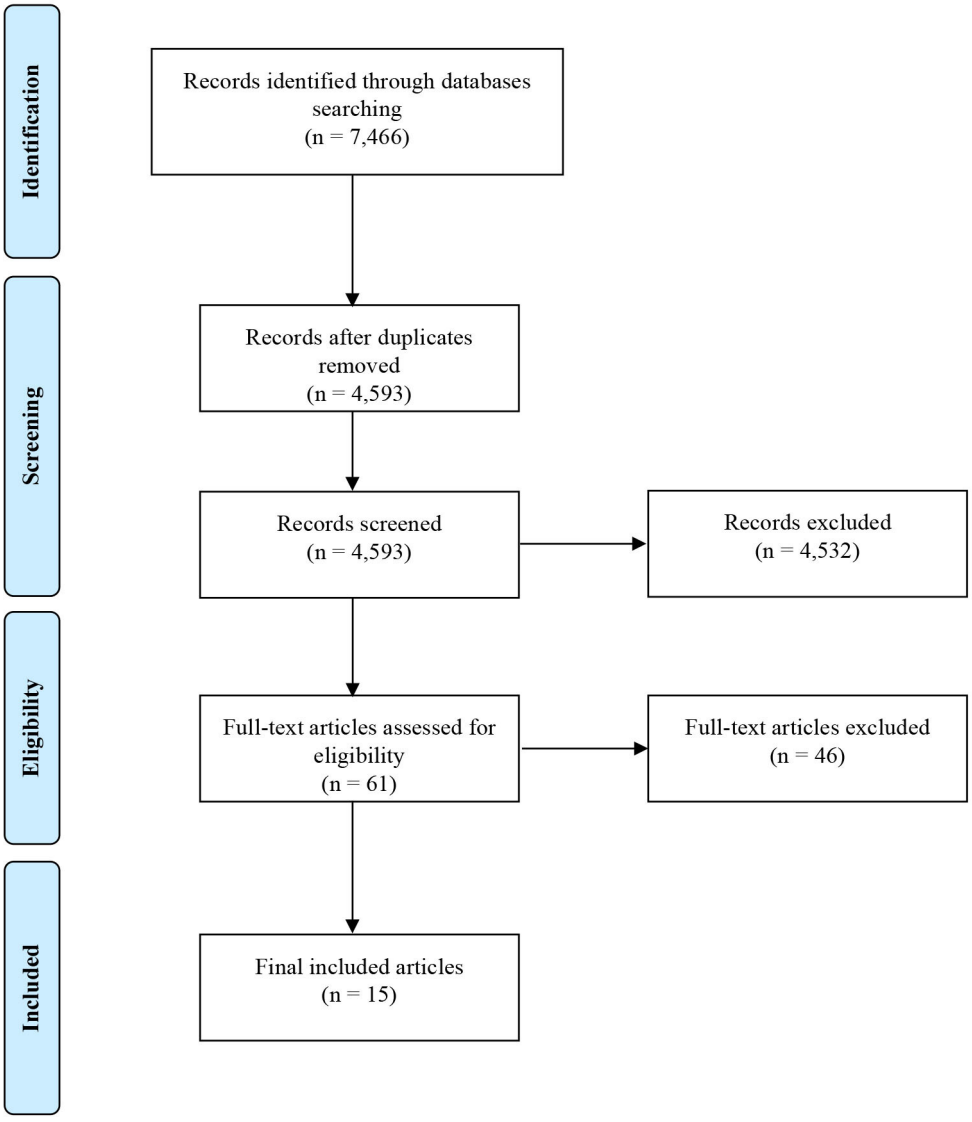

Figure 1 Flow diagram of study selection (PRISMA chart). PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

records in some studies.[21 29] Different approaches, including replacing missing values with median,[33 39] considering missing values as a special value[38] and imputation using GP technique[40] were applied to impute missing values.

Outcome and time horizon

The percentage of target patients in the nine included studies is less than 6%. It should be noted that Liu *et al*[31] and Karlsson *et al*[32] have conducted their studies on small data sets containing 100 and 445 patients from which 40 and 63 patients died during hospitalisation, respectively. Since the outcome patients in most studies were few, large data sets should be collected to obtain enough event patients.

As the main aim of the included studies was building ML models to predict mortality at EDs, these articles should specify how much earlier their models can predict mortality. However, as shown in table 5, only six studies reported their time horizon for mortality prediction; 9–12 min, 2 hours, 24 hours, 72 hours, 7 days, 28 days, 30 days and 31 days ahead were reported as time horizons. Since studies in this area try to predict mortality in advance, all studies must report their time horizons, meaning the time that their models provide clinicians to intervene before mortality occurs.

Imbalanced data

Fourteen studies utilised classification approaches to predict mortality at EDs. Since the number of patients who died during hospitalisation (table 4) for many studies is small, some proper preprocessing techniques such as balancing classes are needed. In addition, except,[40] all other studies stratified patients into two groups, for example, non-survived patients and survived patients, but the number of patients who died during hospitalisation was much lower. This challenge, which is called imbalanced data, can affect the performance of ML models significantly; thus, proper techniques should be applied to address this challenge.

Table 4 General characteristics of included studies

ID	Authors	Year	Country	Study type	Population	Outcome portion (%)	Patients type	Mean age (y)	Male (%)	Vital signs
A1[18]	Levin *et al*	2017	USA	Retrospective	172 726	0.40	All admissions	46	44.7	TP, HR, RR, SBP, SpO_2
A2[19]	Klug *et al*	2020	Israel	Retrospective	799 522	1.55	All admissions	55	51.5	TP, HR, SBP, DBP, SpO_2
A3[21]	Faisal *et al*	2020	UK	Retrospective	24 696	5.33	All admissions	63	46.9	RR, SpO_2, SBP, PR, GCS, TP
A4[22]	Kwon *et al*	2020	South Korea	Retrospective	23 587	3.98	Infectious patients	63	46.1	SBP, RR, TP, HR
A5[29]	Raita et al	2019	USA	Retrospective	135 470	2.10	All admissions	46	56.8	TP, PR, SBP, DBP, RR, SpO_2
A6[31]	Liu *et al*	2011	Singapore	Retrospective	100	40.00	All admissions	65	63	RR, PR, SBP, DBP, SpO_2, GCS
A7[32]	Karlsson *et al*	2021	Sweden	Retrospective	445	14.1	Infectious patients	73	52.6	HR, SBP, DBP, RR, TP, SpO_2
A8[33]	Chen *et al*	2021	Taiwan	Retrospective	52 626	9.4	Infectious patients	72	58.3	HR, SBP, DBP, TP, RR, SpO_2, GCS
A9[34]	Joseph *et al*	2020	USA	Retrospective	445 925	<13	All admissions	53	45.9	HR, SBP, DBP, RR, TP, SpO_2
A10[35]	Rodriguez *et al*	2021	Columbia	Retrospective/ prospective	2510	11.5	Infectious patients	62	49.8	SBP, DBP, TP, RR, SpO_2, GCS
A11[36]	Soffer *et al*	2020	Israel	Retrospective	118 262	5.3	All admissions	73	52.6	HR, SBP, DBP, TP, RR, SpO_2
A12[37]	van Doorn *et al*	2021	Netherlands	Retrospective	1344	13	Infectious patients	71	54.4	HR, SBP, DBP, TP, RR, SpO_2, GCS
A13[38]	Taylor *et al*	2015	USA	Retrospective	5278	4.92	Infectious patients	65	55	HR, TP, SBP, DBP, RR, SpO_2
A14[39]	Perng *et al*	2019	Taiwan	Retrospective	42 220	4.71	Infectious patients	64	56.5	SBP, GCS, TP, HR
A15[40]	Shamout *et al*	2020	UK	Retrospective	37 284	0.80	All admissions	68	48.8	HR, SBP, RR, TP, SpO_2

DBP, diastolic blood pressure; GCS, Glasgow coma scale; HR, heart rate; PR, pulse rate; RR, respiratory rate; SBP, systolic blood pressure; SpO_2, oxygen saturation; TP, body temperature.

Nevertheless, only two studies[22 33] mentioned this challenge and used Synthetic Minority Over-sampling Technique (SMOTE)[43] to address it.

Machine learning models

As shown in table 5, eight ML techniques have been utilised in the included studies. RF was the most common ML technique and was used in nine studies. A brief description and list of advantages and disadvantages of each ML technique is summarised in table 6. Although most studies have been made during the last 5 years, we observed traditional algorithms such as LR more than advanced algorithms, applied to build predictive models in five studies.

Validation and evaluation

Ten articles used k-fold cross-validation,[21 22 29 31 32 34 35 37–39] while one study applied bootstrapping technique for validation purpose.[36] Two studies split their data into training and test sets and used a portion of the training set for validation.[19 40] Three studies investigated the generalisation capability of their models using external validation, while the rest of the studies were internally validated. Researchers have used different metrics for evaluation of ML models in predicting mortality at EDs. The area under the curve (AUC) metric was used in 14 articles. Eight studies assessed the performance of ML models using sensitivity and specificity. Five studies presented the accuracy of models, while positive predictive value and negative predictive value were used in three studies. The Brier Score, positive likelihood ratio, negative likelihood ratio, and false positive rate were less common metrics used in two studies, one, one, and one study, respectively.

Personalised model

Patients' characteristics should be considered in developing predictive models and the potential impact on the improvement of models' performance should be investigated. However, among the included studies, only one study used information of each patient to predict the risk of mortality for the same patient.[40] They considered patients' vital sign trajectories as time series data and introduced an ML model to follow-up the trends of vital signs for each patient and by that they tried to consider patients' characteristics, while the other studies only build one model for the whole population and categorised patients into two groups and binary classifiers using different ML algorithms were built.

Table 5 Selected articles list and their ML-related characteristics

Id	ML algorithms	Evaluation metrics	Time horizon	Personalised	Handling missing values	Hyperparameter opgtimisation	Approach	Validation
A1	RF	AUC	–	No	No	No	Binary classification	Internal
A2	GB	AUC, sensitivity, specificity, Brier Score	–	No	No	No	Binary classification	Internal
A3	LR, RF, SVM, NN	AUC, Brier Score	–	No	No	Yes	Binary classification	External
A4	GB, RF	AUC	72 hours	No	No	Yes	Binary classification	External
A5	LR, RF, GB DT, DNN	AUC, sensitivity, specificity	–	No	No	Yes	Binary classification	Internal
A6	NN, SVM	Accuracy, sensitivity, specificity	9–12 min	No	No	No	Binary classification	Internal
A7	RF	AUC, sensitivity, specificity, PPV, NPV, PLR, NLR	7 days, 30 days	No	No	No	Binary classification	Internal
A8	SVM, GB, NN	AUC, accuracy, sensitivity, specificity, PPV, NPV	–	No	Yes	No	Binary classification	Internal
A9	LR, DNN, GB	AUC, accuracy, sensitivity, specificity,	–	No	No	Yes	Binary classification	Internal
A10	DT, RF, NN, SVM	AUC, accuracy	–	No	No	No	Binary classification	Internal
A11	GB	AUC, sensitivity, specificity, NPV, PPV, FPR	–	No	Yes	No	Binary classification	Internal
A12	LR, NN, RF, GB	AUC, sensitivity, specificity	31 days	No	No	No	Binary classification	Internal
A13	RF, DT, LR	AUC	–	No	Yes	No	Binary classification	Internal
A14	KNN, SVM, RF, DNN	AUC, accuracy	72 hours, 28 days	No	Yes	No	Binary classification	Internal
A15	DNN	AUC	2 hours, 24 hours	Yes	Yes	No	Time series regression	External

AUC, area under the curve; DNN, deep neural networks; DT, decision tree; FPR, false positive rate; GB, gradient boosting; KNN, K-nearest neighbours; LR, logistic regression; ML, machine learning; NLR, negative likelihood ratio; NN, neural networks; NPV, negative predictive value; PLR, positive likelihood ratio; PPV, positive predictive value; RF, random forest; SVM, support vector machine.

Risk of bias assessment

There is a significant amount of incomplete reporting within the studies' results. For instance, most studies did not report information about the tuning process of ML models and hyperparameter values considered during the model training process. A summary of the PROBAST assessment is shown in table 7. As shown, participants' selection was done properly except in two studies. However, most studies had a high ROB in the statistical analysis domain. Among the studies, only Joseph *et al*[34] utilised transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)[44] guideline and provide adequate information about feeding data to models and architecture of implemented NN, so only this study was considered as low ROB in the statistical domain.

DISCUSSION

Weaknesses and strengths of included studies

This review consisted of 15 studies that developed ML-based models to predict mortality in EDs. Our analysis showed that most studies described the data sets in sufficient detail and had enough number of participants. Moreover, the outcome domain had the lowest ROB which shows most studies had a clear outcome definition. However, we found several methodological and reporting shortcomings, and this can lead to poorer performance than reported when the models are deployed in real life. Many studies skipped or failed to report important phases such as data preprocessing, handling missing values and model development. It has been stated that up to 80% of data analysis is spent on data cleaning and

Table 6 Summary of machine learning algorithms' description, advantages and disadvantages

Algorithm	Description	Advantages	Disadvantages
LR[58]	LR is a supervised ML algorithm adopted from linear regression. It can be used for classification problems and finding the probability of an event happening.	Fast training, good for small data sets, easy to understand.	Not very accurate, not proper for non-linear problems, high chance of overfitting, not flexible to adopt to complex data sets.
DT[59]	DT is a supervised ML algorithm that solves a problem by transforming the data into a tree representation where each internal node represents an attribute, and each leaf denotes a class label.	Easy to understand and interpret, robust to outliers, no standardisation or normalisation required, useful for regression and classification.	High chance of overfitting, not suitable for large data sets, adding new samples lead to regeneration of the whole tree.
KNN[60]	KNN is a supervised and instance-based ML algorithm. It can be used when we want to forecast a label of a new sample based on similar samples with known labels. Different similarity or distance measures such as Euclidean can be used.	Simple and easy to understand, easy to implement, no need for training, useful for regression and classification.	Memory intensive, costly, slow performance, all training data might be involved in decision-making.
SVM[61]	SVM is an instance-based and supervised ML technique that generates a boundary between classes known as hyperplane. Maximising the margin between classes is the main goal of this technique.	Efficient in high dimensional spaces. Effective when the number of dimensions is greater than the number of samples, long training time, useful for regression and classification.	Not suitable for large data sets, not suitable for noisy data sets, Regularisation capabilities which prevent overfitting, handling non-linear data.
GB[62]	GB is a supervised ML algorithm, which produces a model in the form of an ensemble of weak prediction models, usually DT. GB is an iterative gradient technique that minimises a loss function by iteratively selecting a function that points towards the negative gradient.	High accuracy, high flexibility, fast execution, useful for regression and classification, robust to missing values and overfitting.	Sensitive to outliers, not suitable for small data sets, many parameters to optimise.
RF[63]	RF is an ensemble and supervised ML algorithm that is based on the bagging technique, which means that many subsets of data are randomly selected with replacement and each model such as DT is trained using one subset. The output is the average of all predictions of various single models.	High accuracy, fast execution, useful for regression and classification, robust to missing values and overfitting.	Not suitable for limited data sets, may change considerably by a small change in the data.
NN[64]	NN is a family of supervised ML algorithms. It is inspired by biological neural network of the human brain. NN consists of input, hidden, output layers and multiple neurons (nodes) carry data from input layer to output layer.	Accurate, suitable for complex, non-linear classification and regression problems.	Very slow to train and test, requires large amount of data, computationally expensive and prone to overfitting.
DNN[65]	DNN is a family of supervised ML algorithms. DNN is based on NN where the adjective 'deep' comes from the use of multiple layers in the network. Usually having two or more hidden layers counts as DNN. There are some specific training algorithms and architecture such as LSTM, GAN, CNN for DNNs. DNNs provide the opportunity to solve complex problems when the data are very diverse, unstructured and interconnected.	High accuracy, features are automatically deduced and optimally tuned, robust to noise, architecture is flexible.	Needs very large amount of data, computationally expensive, not easy to understand, no standard theory in selecting the right settings, difficult for less skilled researchers.

CNN, convolutional neural networks; DNN, deep neural networks; DT, decision tree; GAN, generative adversarial networks; GB, gradient boosting; KNN, K-nearest neighbours; LR, logistic regression; LSTM, long-short term memory networks; ML, machine learning; NN, neural networks; RF, random forest; SVM, support vector machine.

data preparation,[41] so these phases are very crucial for having an ML predictive model with good performance. It also could be one of the reasons why some researchers have discussed the poor effectiveness of ML predictive models in clinical environments.[45 46] Moreover, despite the dramatic advances in ML and introducing novel and effective models, various researchers still have applied traditional ML techniques such as LR and DT. For instance, we found that LR is the third most common algorithm used in the studies. Most ML models have parameters that needed to be tuned to gain the best performance and generalisation power and avoiding overfitting, but most studies did not provide information about the models' validation process and finding the hyperparameters of the models. Of 15 studies, 14 studies

Table 7 Risk of bias (ROB) assessment of included studies according to PROBAST checklist

	PROBAST items				
ID	Participants	Predictors	Outcomes	Sample size and missing data	Statistical analysis
A1	+	+	+	–	–
A2	+	?	+	–	–
A3	+	+	+	–	–
A4	+	?	+	–	–
A5	+	+	+	–	–
A6	–	?	+	–	–
A7	–	+	+	–	–
A8	+	+	+	+	–
A9	+	+	+	?	+
A10	?	+	+	–	?
A11	+	+	+	?	?
A12	?	+	+	–	?
A13	+	+	?	+	?
A14	+	?	?	–	?
A15	+	+	+	+	?

+, low risk of bias; –, high risk of bias; ?, unclear risk of bias.
PROBAST, prediction risk of bias assessment tool.

are retrospective studies and only three studies have evaluated their models based on external validation.

The ultimate aim of included studies was to predict mortality in EDs ahead of time to provide enough time for clinicians to intervene and prevent mortality. However, only six studies have provided their time horizon, the time that clinicians have to treat the patients. The other studies have built binary classification algorithms to stratify patients into two classes, and their models do not have the capability to be used in practice.

Clinical considerations and prospective

A recent systematic review on the value of vital sign trends in monitoring and predicting clinical deterioration showed a lack of research and knowledge regarding the importance of vital signs in clinical deterioration of ED patients.[9] Even though considering the vital sign as cheap and available clinical predictors, there is a hypothesis stating the lack of evidence to support the usefulness of continuous monitoring of vital signs as a daily routine of clinical practice.[47–49] The other barrier refers to the comparison between ML methods and conventional LR for clinical prediction models. A recent systematic review showed no performance benefits of ML methods over LR.[50] However, as mentioned in the previous section, this conclusion could be due to lack of enough attention to the important stages and practical issues such as data preparation or models' tuning, and consequently results are not satisfactory and implemented models are not robust. Therefore, further studies need to be performed to prove the superiority of ML algorithms over conventional models. It seems that there is still a long way to go for having ML algorithms as the choice of clinical deterioration in EDs. The increase of studies in this field during the last few years proves an increased attention to appraise the capabilities of ML methods in clinical practice.

Challenges and recommendations for future work

Perhaps in this research area, the main challenge is integration of ML models into clinical practice. A recent systematic review based on performance metrics, particularly AUC, showed that ML predictive models outperformed usual care in most detection and prediction tasks at ED. However, the authors mentioned that many studies have limited applicability to clinical practice and there are other considerations more than performance metrics as well as barriers that should be taken into account to have successful ML models in real life.[51] Therefore, despite the great research successes in building ML-based predictive models for clinical practice, there remain few examples of ML models being successfully integrated into the daily routine or critical parts of clinical environments.[52] This reveals that what is being done in research is not completely in line with the realities of clinical practice. In this section, we mention some of these issues and technical barriers, then present recommendations on how to address the most prominent issues and encourage researchers to take them into account.

Our analysis indicated that most studies are retrospective. However, the impact of ML techniques in clinical environment will need further validation in randomised

control trials and prospective studies before widespread clinical adoption where the data could be incomplete and noisy with high level of uncertainty. *Recommendation 1*: we recommend that besides building models using historical collected data, researchers consider prospective analysis and think about building adaptive models that can monitor patients in real-time.

Although external validation is one of the most rigorous ways to assess the generalisation power of predictive models, it is still one of the common approaches for showing the robustness of predictive models.[27 51] *Recommendation 2*: we recommend that researchers investigate the performance of their models using external validation to have a better estimation of the generalisation power of their models. *Recommendation 3*: regarding the preprocessing phase, since some steps such as data cleaning, handling missing and noisy values are an inevitable part of each data analysis study, we recommend that researchers instead of ignoring records with missing values, use proper techniques to handle noisy and incomplete records. For instance, we recommend that complete case analysis is avoided and proper imputation techniques such as iterative multivariate imputation technique,[53] which is one of the best approaches for imputing missing values is used instead.[54] Another challenge in most studies is that the number of event patients is too low compared with non-event patients. This can lead to bias and overfitting during developing ML models, so proper under/oversampling techniques such as SMOTE and its variations should be utilised to make a balance between classes before developing models.

Another issue is time horizon for prediction. Gerry *et al* recommended that time horizon should be limited to a few days at most, since signs of deterioration will probably not be observed for more than a few days.[55] *Recommendation 4*: we suggest that researchers in this area choose a realistic time horizon rather than prediction of mortality after 6 or 12 months. Moreover, most of the included studies are based on building a single model for the whole population. *Recommendation 5*: as patients have different characteristics and norms, we recommend that patients' characteristics, vital signs trajectories, trends, etc, are used for each individual patient to tune ML models to have a better performance in practice. *Recommendation 6*: we recommend that researchers find a way to make their results and workflow understandable and interpretable by humans. For instance, deep learning is one of the hottest research topics today, and it has been applied in different applications of clinical practice, but it has inherent limitations. For example, despite impressive results, it is not indicative of high-level reasoning and rationality. Such methods are often considered as a 'black box', where the decision logic is presented in millions of numerical weights and biases. This lack of transparency could have some legal and ethical implications and might increase distrust of ML models by patients and clinicians, and it might be one of the reasons that patients are likely to trust a clinician more than a machine. This has been described as a crucial problem for ML acceptance in clinical practice and attempts are underway to create more human interpretable models. For example, with the help of image-processing techniques, it is possible now to visualise the input and output of ML models; this makes the understanding of ML models easier.[17]

Lack of widely accepted and utilised reporting or publication guidelines for implementing ML in clinical practice is a challenge that brings difficulty to the research quality assessment, especially for clinicians without a strong mathematics and computer science background.[17] Another challenge for ML models is that they need a huge amount of labelled data to be trained, and the performance of ML models often depends on the performance of the human labelling the data. *Recommendation 7*: as this area is interdisciplinary in nature, it is imperative that clinicians and artificial intelligence experts become better at interdisciplinary collaboration to become more familiar with the needs, limitations and challenges in clinical and technical domains, which ultimately increases the quality of the predictive models. *Recommendation 8*: we recommend that researchers should be careful and accurate when reporting their studies. It has been shown that the quality of reporting of predictive models in this clinical domain is poor. Only with full and transparent reporting of information about all aspects of a predictive model, the ROB and the usefulness of that model can be assessed. It is recommended that researchers use available checklists such as Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence (SPIRIT-AI)[56] for designing the trial, and TRIPOD[44] and Consolidated Standards of Reporting Trials - Artificial Intelligence (CONSORT-AI)[57] for reporting the findings to make sure that they provide sufficient information to clearly and accurately report their studies.

Strengths and limitations of this study

To the best of our knowledge, this study is the most recent systematic review on the topic of using ML algorithms to predict mortality at EDs. This study was done with the help of an experienced artificial intelligence expert, an experienced research librarian who is familiar with the health research area, and three expert clinicians at ED. Moreover, ROB analysis was done using PROBAST checklist which has four domains that assess different aspects of studies. Data extraction was done using a standard checklist called CHARMS. However, our study has some limitations. First, it was not possible to quantify the analysis or a conduct meta-analysis due to the high heterogeneity of included studies. Also, we have considered the vital signs due to high availability at ED, while the role of other clinical predictors in patients' clinical conditions should be investigated.

CONCLUSION

The application of ML methods to identify clinical deterioration remains equally challenging as identification of deterioration using track and trigger protocols and similar human-driven protocols. However, since the ML approaches are as diverse as the problem they deal with, assessment of various methods and their performance are needed. This systematic review was performed on the topic of utilisation of ML techniques to predict mortality at EDs using vital signs. Our systematic review of the literature provides an updated overview of the state-of-the-art on this topic (covering 2010 to 2021). Initially, 7466 potential articles were identified of which 15 were included in the analysis. After a comprehensive review of these 15 articles, we generated eight recommendations to increase the feasibility of implementing ML models in EDs. These recommendations provide actionable suggestions to be used to increase the quality of future work in this area.

Acknowledgements The authors would like to thank Dr Mette Brandt Eriksen, research librarian at University of Southern Denmark library for help with the search strategy and compiling the literature databases, and Professor Annmarie Touborg Lassen, Professor John Kellet and Professor Mikkel Brabrand, expert emergency department clinicians at Odense University Hospital, for their help during performing this systematic review.

Contributors All authors were involved in conception and design of the study. Literature search was conducted by AN under supervision of an expert librarian. AN, TS, MM and AE participated in screening and extracting relevant studies. AN, MN-B, TS and UKW were responsible for analysing, and drafting the manuscript. All authors made critical revisions. All authors approved the final manuscript. The corresponding author (AN) takes responsibility for the study as a whole.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Formal ethical approval is not required, as primary data will not be collected. The review will be disseminated in peer-reviewed publications.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. All data relevant to the study are included in the article or uploaded as supplementary information. Moreover, Spreadsheets can be accessed via the Dryad data repository at https://doi.org/10.5061/dryad.8w9ghx3nc.

ORCID iDs

Amin Naemi http://orcid.org/0000-0003-4501-4310
Thomas Schmidt http://orcid.org/0000-0003-4476-8559
Marjan Mansourvar http://orcid.org/0000-0001-6492-7858
Mohammad Naghavi-Behzad http://orcid.org/0000-0002-6761-8126
Ali Ebrahimi http://orcid.org/0000-0002-3332-6205
Uffe Kock Wiil http://orcid.org/0000-0001-6898-4083

REFERENCES

1 Leon DA, Jdanov DA, Shkolnikov VM. Trends in life expectancy and age-specific mortality in England and Wales, 1970-2016, in comparison with a set of 22 high-income countries: an analysis of vital statistics data. *Lancet Public Health* 2019;4:e575–82.
2 Preston L, van Oppen JD, Conroy SP, *et al*. Improving outcomes for older people in the emergency department: a review of reviews. *Emerg Med J* 2020 doi:10.1136/emermed-2020-209514
3 Conway R, Byrne D, O'Riordan D, *et al*. Comparative influence of acute illness severity and comorbidity on mortality. *Eur J Intern Med* 2020;72:42–6.
4 Wretborn J, Khoshnood A, Wieloch M, *et al*. Skåne emergency department assessment of patient load (SEAL)-a model to estimate crowding based on workload in Swedish emergency departments. *PLoS One* 2015;10:e0130020.
5 Hasler RM, Albrecht S, Exadaktylos AK, *et al*. Repatriations and 28-day mortality of ill and injured travellers: 12 years of experience in a Swiss emergency department. *Swiss Med Wkly* 2015;145:w14208.
6 Hall MJ, Levant S, DeFrances CJ. *Trends in inpatient hospital deaths: National hospital discharge survey, 2000-2010*. US Department of Health and Human Services, Centers for Disease Control, 2013.
7 Jeffery AD, Dietrich MS, Fabbri D, *et al*. Advancing in-hospital clinical deterioration prediction models. *Am J Crit Care* 2018;27:381–91.
8 Escobar GJ, Liu VX, Schuler A, *et al*. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med* 2020;383:1951–60.
9 Brekke IJ, Puntervoll LH, Pedersen PB, *et al*. The value of vital sign trends in predicting and monitoring clinical deterioration: a systematic review. *PLoS One* 2019;14:e0210875.
10 Kellett J. The assessment and interpretation of vital signs. In: *Textbook of rapid response systems*. Springer, 2017: 63–85.
11 Henriksen DP, Brabrand M, Lassen AT. Prognosis and risk factors for deterioration in patients admitted to a medical emergency department. *PLoS One* 2014;9:e94649.
12 Barfod C, Lauritzen M, Danker J, *et al*. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department - a prospective cohort study. *Scand J Trauma Resusc Emerg Med* 2012;20:28–10.
13 Buist M, Bernard S, Nguyen TV, *et al*. Association between clinically abnormal observations and subsequent in-hospital mortality: a prospective study. *Resuscitation* 2004;62:137–41.
14 Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;349:255–60.
15 Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
16 Sathya R, Abraham A. Comparison of supervised and unsupervised learning algorithms for pattern classification. *Int J Adv Res Art Intel* 2013;2:34–8.
17 Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *Emerg Med Australas* 2018;30:870–4.
18 Levin S, Toerper M, Hamrock E, *et al*. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Ann Emerg Med* 2018;71:565–74.
19 Klug M, Barash Y, Bechler S, *et al*. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a Nine-Point triage score. *J Gen Intern Med* 2020;35:220–7.
20 Dugas AF, Kirsch TD, Toerper M, *et al*. An electronic emergency triage system to improve patient distribution by critical outcomes. *J Emerg Med* 2016;50:910–8.
21 Faisal M, Scally A, Howes R, *et al*. A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. *Health Informatics J* 2020;26:34–44.
22 Kwon YS, Baek MS. Development and validation of a quick sepsis-related organ failure assessment-based machine-learning model for mortality prediction in patients with suspected infection in the emergency department. *J Clin Med* 2020;9:875.

23 Ye C, Wang O, Liu M, *et al*. A real-time early warning system for monitoring inpatient mortality risk: prospective study using electronic medical record data. *J Med Internet Res* 2019;21:e13719.
24 Brajer N, Cozzi B, Gao M, *et al*. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. *JAMA Netw Open* 2020;3:e1920733.
25 Kaieski N, da Costa CA, da Rosa Righi R. Application of artificial intelligence methods in vital signs analysis of hospitalized patients: a systematic literature review. *Applied Soft Computing* 2020;106612.
26 Moons KGM, de Groot JAH, Bouwmeester W, *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744.
27 Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51–8.
28 Debray TPA, Damen JAAG, Snell KIE, *et al*. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
29 Raita Y, Goto T, Faridi MK, *et al*. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019;23:64.
30 Riley RD, Snell KI, Ensor J, *et al*. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38:1276–96.
31 Liu N, Lin Z, Koh Z, *et al*. Patient outcome prediction with heart rate variability and vital signs. *J Signal Process Syst* 2011;64:265–78.
32 Karlsson A, Stassen W, Loutfi A, *et al*. Predicting mortality among septic patients presenting to the emergency department-a cross sectional analysis using machine learning. *BMC Emerg Med* 2021;21:84.
33 Chen Y-M, Kao Y, Hsu C-C, *et al*. Real-time interactive artificial intelligence of things-based prediction for adverse outcomes in adult patients with pneumonia in the emergency department. *Acad Emerg Med* 2021 doi:10.1111/acem.14339
34 Joseph JW, Leventhal EL, Grossestreuer AV, *et al*. Deep-learning approaches to identify critically ill patients at emergency department triage using limited information. *J Am Coll Emerg Physicians Open* 2020;1:773–81.
35 Rodríguez A, Mendoza D, Ascuntar J, *et al*. Supervised classification techniques for prediction of mortality in adult patients with sepsis. *Am J Emerg Med* 2021;45:392–7.
36 Soffer S, Klang E, Barash Y, *et al*. Predicting in-hospital mortality at admission to the medical ward: a big-data machine learning model. *Am J Med* 2021;134:227–34.
37 van Doorn WPTM, Stassen PM, Borggreve HF, *et al*. A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS One* 2021;16:e0245157.
38 Taylor RA, Pare JR, Venkatesh AK, *et al*. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016;23:269–78.
39 Perng J-W, Kao I-H, Kung C-T, *et al*. Mortality prediction of septic patients in the emergency department based on machine learning. *J Clin Med* 2019;8:1906.
40 Shamout FE, Zhu T, Sharma P, *et al*. Deep interpretable early warning system for the detection of clinical deterioration. *IEEE J Biomed Health Inform* 2020;24:437–46.
41 García S, Luengo J, Herrera F. *Data preprocessing in data mining*. Springer, 2015.
42 Dziura JD, Post LA, Zhao Q, *et al*. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med* 2013;86:343.
43 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321–57.
44 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD group. *Circulation* 2015;131:211–9.
45 Smith MEB, Chiovaro JC, O'Neil M, *et al*. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Ann Am Thorac Soc* 2014;11:1454–65.
46 Alam N, Hobbelink EL, van Tienhoven AJ, *et al*. The impact of the use of the early warning score (EWS) on patient outcomes: a systematic review. *Resuscitation* 2014;85:587–94.
47 Cardona-Morrell M, Prgomet M, Turner RM, *et al*. Effectiveness of continuous or intermittent vital signs monitoring in preventing adverse events on general wards: a systematic review and meta-analysis. *Int J Clin Pract* 2016;70:806–24.
48 Downey CL, Chapman S, Randell R, *et al*. The impact of continuous versus intermittent vital signs monitoring in hospitals: a systematic review and narrative synthesis. *Int J Nurs Stud* 2018;84:19–27.
49 van Loon K, van Zaane B, Bosch EJ, *et al*. Non-invasive continuous respiratory monitoring on general hospital wards: a systematic review. *PLoS One* 2015;10:e0144626.
50 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.
51 Kareemi H, Vaillancourt C, Rosenberg H, *et al*. Machine learning versus usual care for diagnostic and prognostic prediction in the emergency department: a systematic review. *Acad Emerg Med* 2021;28:184–96.
52 Coiera E. The last mile: where artificial intelligence meets reality. *J Med Internet Res* 2019;21:e16323.
53 Little RJA, Rubin DB. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.
54 Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? *Int J Epidemiol* 2014;43:1336–9.
55 Gerry S, Bonnici T, Birks J, *et al*. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020;369:m1501.
56 Rivera SC, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;370:m3210.
57 Liu X, Rivera SC, Moher D. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020;370.
58 Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.
59 Song Y-Y, Lu Y, Ying LU. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* 2015;27:130.
60 Kramer O. *Dimensionality reduction with unsupervised nearest neighbors*. Springer, 2013.
61 Vapnik V. *The nature of statistical learning theory*. Springer science & business media, 2013.
62 Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot* 2013;7:21.
63 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
64 Abiodun OI, Jantan A, Omolara AE, *et al*. State-of-the-art in artificial neural network applications: a survey. *Heliyon* 2018;4:e00938.
65 Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access* 2019;7:53040–65.