Annals of Clinical Epidemiology

SEMINAR

# Introduction to supervised machine learning in clinical epidemiology

Sachiko Ono[1], Tadahiro Goto[2,3]

[1] *Department of Eat-loss Medicine, Graduate School of Medicine, The University of Tokyo*

[2] *Department of Clinical Epidemiology and Health Economics, The University of Tokyo*

[3] *TXP Medical Co. Ltd.*

**ABSTRACT**

Machine learning refers to a series of processes in which a computer finds rules from a vast amount of data. With recent advances in computer technology and the availability of a wide variety of health data, machine learning has rapidly developed and been applied in medical research. Currently, there are three types of machine learning: supervised, unsupervised, and reinforcement learning. In medical research, supervised learning is commonly used for diagnoses and prognoses, while unsupervised learning is used for phenotyping a disease, and reinforcement learning for maximizing favorable results, such as optimization of total patients' waiting time in the emergency department. The present article focuses on the concept and application of supervised learning in medicine, the most commonly used machine learning approach in medicine, and provides a brief explanation of four algorithms widely used for prediction (random forests, gradient-boosted decision tree, support vector machine, and neural network). Among these algorithms, the neural network has further developed into deep learning algorithms to solve more complex tasks. Along with simple classification problems, deep learning is commonly used to process medical imaging, such as retinal fundus photographs for diabetic retinopathy diagnosis. Although machine learning can bring new insights into medicine by processing a vast amount of data that are often beyond human capacity, algorithms can also fail when domain knowledge is neglected. The combination of algorithms and human cognitive ability is a key to the successful application of machine learning in medicine.

KEY WORDS

supervised learning, random forests, gradient-boosted decision tree, support vector machine, neural network

## 1. INTRODUCTION

The availability of various health care data, including electronic health records, registries, claims data, and digital imaging, has been proliferating in the past few decades. These data are linked, integrated, and utilized for medical research [1]. Recent advances in computer technology and statistics enabled the implementation of complex and computationally expensive algorithms with large-scale data. With the combination of these data and technological advancements, researchers have attempted to develop algorithms—termed machine learning—that imitate and even excel in human cognitive ability to do complex tasks in the medical field.

Machine learning refers to an algorithm in which a computer recognizes patterns and relationships of variables based on given data. Each algorithm develops a model to output an answer for a specific problem. Researchers have tried to develop machine learning algorithms that substitute for medical experts, or that find unexpected rules that are beyond human comprehension. These attempts are prompted by the facts that advancement of medical science have invented multiple treatment options and more subdivided diagnosis for a disease, resulting in ever more complex decision-making process in practice and shortage of experts of such subcategories. In this context, machine learning has been increasingly utilized in medical research, leveraging abundant medical data. The number of articles in PubMed that are tagged with "machine learning" [Mesh] has increased dramatically since the MeSH term introduction, as shown in **Fig. 1**.

Compared with the conventional regression model and prediction method, machine learning can handle more variables and build a complex model that considers interactions of variables and nonlinear relationships between variables and outcomes. Currently, there are three types of machine learning: supervised, unsupervised, and reinforcement learning. The present article focuses on supervised learning, a type of machine learning commonly used for the prediction and diagnoses problems, introducing the concept, commonly used four algorithms, and their applications in medical research.

## 2. CONCEPT OF MACHINE LEARNING

Machine learning refers to a series of processes in which a computer finds rules from a vast amount of data. In machine learning, the computer develops a model that represents what it has learned from the data (i.e., the relationships among variables), and applies the model to unknown data to make predictions and classifications. The development of machine learning has been driven by the development of databases in recent years. The digitization of various documents and automatic recording by electronic sensors have enabled the constant collection of vast amounts of data. Such data include medical claims, electronic health records, laboratory data, and medical images from various medical fields [1].

In a conventional model such as logistic regression model, a human determines variables (also referred to as predictors in a prediction model), and develops a model based on domain knowledge to predict outcomes. (To be precise, logistic regression analysis can also be categorized as machine learning due to its iterative process for maximum likelihood estimation, but the term "machine learning" is rarely used for logistic regression analysis in medical articles.) Developing a conventional model is far more difficult when there are enormous number of variables; nonetheless, leveraging such data can bring new insights on a given topic. Machine learning replaces most of the model-creating work with computer algorithms to process vast amounts of data beyond human capacity. Algorithms developed by machine learning methods are particularly powerful when the problem is complex, that is, when the number of variables is huge, when the variables have complex interactions or effect modifiers, and when the association of outcome with variables are nonlinear. Indeed, several studies have shown that machine learning outperforms existing predictive models for complex classification problems such as predicting prognosis and diagnosis. For example, Tokodi et al. [2] developed a machine learning-based risk stratification model to estimate 1 to 5-year mortality risk for patients undergoing cardiac resynchronization therapy, and the model had a much higher predictive ability than all the pre-existing scoring systems (Seattle Heart Failure Model, VALID-CRT, EAARN, ScREEN, and CRT-score).

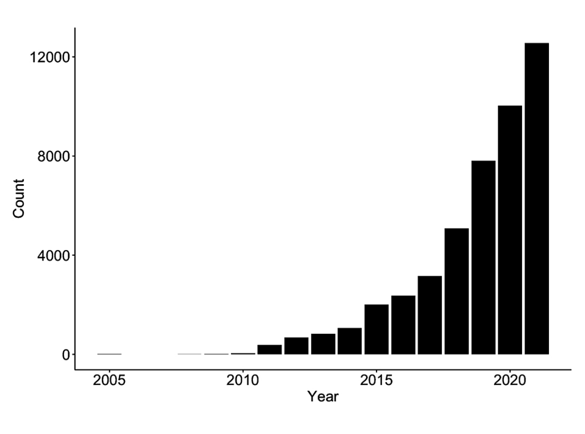

**Fig. 1 Search results of "Machine Learning" [Mesh] by year in PubMed**

### 2.1 Supervised Learning, Unsupervised Learning, and Reinforcement Learning

Conventionally, there are three types of machine learning: supervised, unsupervised, and reinforcement learning. Supervised learning is a type of machine learning in which machines learn from "labeled" training data and then predict the outcome, specifically diagnosis and prognosis in medical field. "Labeled" means that the training data are tagged with the correct answer (i.e., outcome). Unsupervised learning, on the other hand, classifies data with similar characteristics or patterns into groups based on unlabeled data. For example, Bleecker et al. [3] divided patients with asthma into six clinical phenotypes using an unsupervised learning method, mainly for future investigation of pathology and treatment response. In the reinforcement learning, algorithms learn from trial and error (i.e., rewarding desirable results and punishing unwanted ones) to maximize favorable results in cases where there is no given right answer. Lee et al. [4], for example, successfully minimized patients' waiting time in emergency department using reinforcement learning. Among these three machine learning methods, supervised learning is the most frequently utilized method in medical research.

### 2.2 Supervised Learning Algorithms

Although there is a myriad of algorithms in supervised learning, the ones frequently used in medical papers are: (i) random forests, (ii) gradient-boosted decision tree (GBDT) (iii) support vector machines (SVM), and (iv) neural networks [5]. These algorithms can be implemented using statistical software, such as the *caret* package of the R statistical software [6, 7] and the scikit-learn library of Python [8], which effectively help develop machine learning models.

*Decision tree and random forest*

To explain random forests, we will first explain decision trees. A decision tree is a algorithms of determining the final classification by creating branches from each step (node) based on given rules. **Fig. 2** shows an example of a decision tree for deciding whether the rent of a property is 1,000 USD or more. Each rectangle represents a node, of which the first node is called the "root node" and the last node is called the "terminal node". When splitting a node, the cluttering (impurity) of the node's data content is expressed as a numerical value called entropy. The node is split to organize its content; that is, the ratio or difference (information gain) of the entropy of the nodes before and after is maximized [9, 10] (**Fig. 3**). Along with entropy, Gini coefficient is also commonly used to measure the impurity value of a split condition [11].

As the tree grows downward, the data can be subdivided according to its characteristics. While complex data can be finely classified, the inherent noise of the data is also captured and classified as a feature. This may cause overfitting; a model that has learned the noise of one particular dataset will not apply well to another new dataset [12]. This is where "random forests" comes in. "Random forests" is a type of ensemble learning (general term for algorithms combining multiple models) that seeks better predictive performance for new data [13]. Random forests create several decision trees and predict by majority vote to prevent overfitting due to data-specific noise (**Fig. 4**).

In random forests, multiple decision trees are created in parallel using randomly selected data for each. Similarly, the variables used to split the decision tree are also chosen at random (**Fig. 5**). (The variables used in the splitting process are sometimes referred to as "features".) The name "random forests" is derived from the fact that multiple decision trees, that is, "forests", are created using "random" data and variables. Random forests can handle complex problems that cannot be represented by linear models at a relatively high speed.

*Gradient-boosted decision tree (GBDT)*

GBDT is another type of ensemble learning created

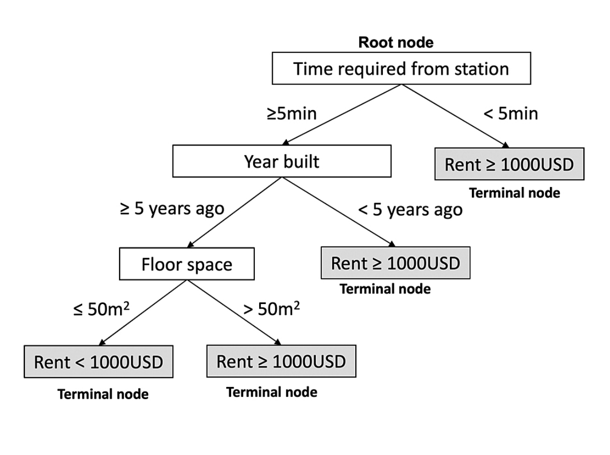

**Fig. 2** An example of a decision tree

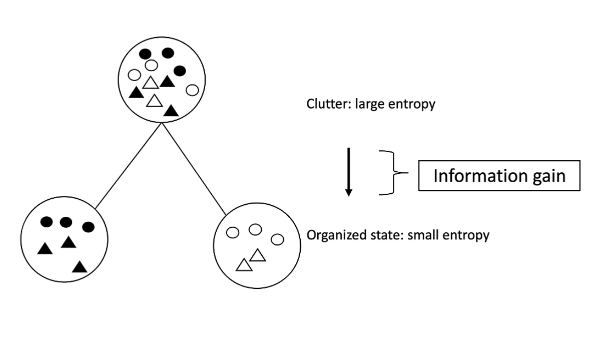

**Fig. 3** Entropy and information gain

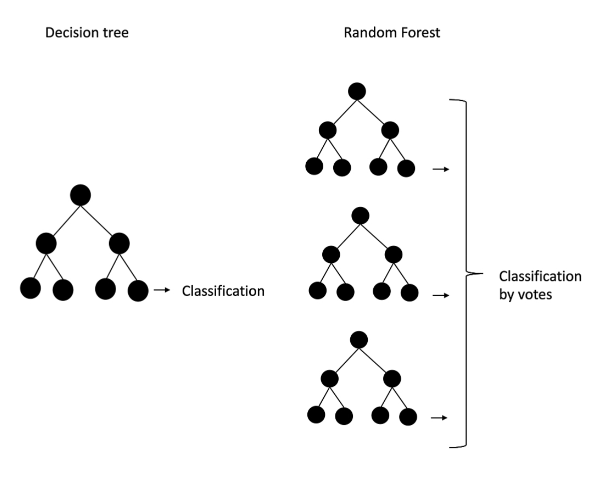

**Fig. 4** Decision tree and random forests

by multiple decision trees. While random forests use multiple decision trees created in parallel, GBDT uses multiple decision trees created in sequence [14]. In this algorithm, the first decision tree produces an initial prediction. The second tree predicts the residuals or errors of the first tree by creating another decision tree. Then, the next model predicts the residuals of the precedent model by creating another decision tree, and this process is repeated until the residuals converge to 0 or the number of iterations reaches a prespecified number (i.e., number of decision trees). After the iterations are done, all trees are combined to make a final prediction (**Fig. 6**). GBDT often outperforms other algorithms in terms of accuracy; however, it may overfit when the number of trees is too large [14].

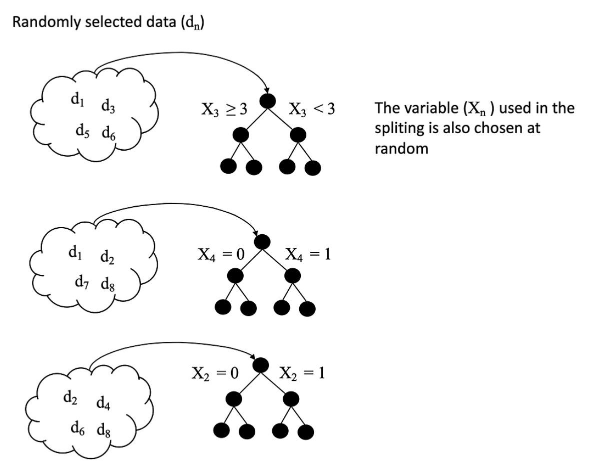

**Fig. 5 Creating random forests**

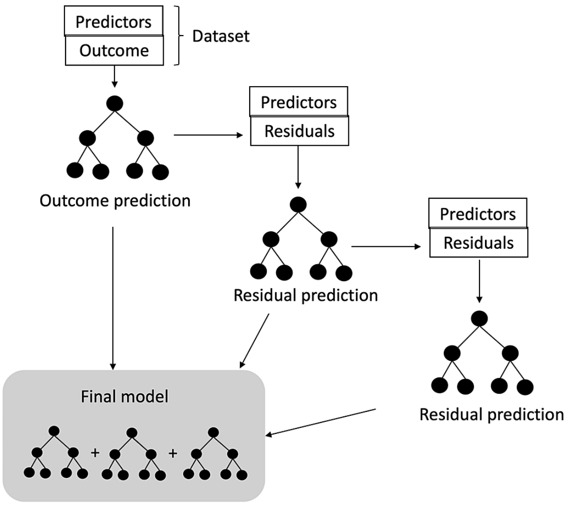

**Fig. 6 Creating gradient-boosted decision tree**

*Support vector machine (SVM)*

SVM classifies data according to a boundary created by an algorithm based on the values of given variables [15–18]. The model developed by the SVM provide a prediction as to which side of the boundary new data will be on. A clinical application is, for example, to develop a model that predicts death or survival within 30 days based on multiple laboratory test results in a certain disease. **Fig. 7** illustrates the use of SVM to draw a boundary line that classifies ● and × based on the values of variables X1 and X2 (e.g., laboratory test results). When creating the boundary, the sum of the distances from the boundary to the data should be maximized.

When data points cannot be separated by a linear boundary, as shown in **Fig. 8**, a method called kernel

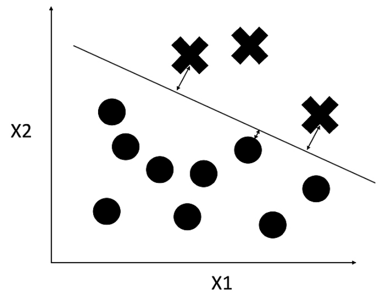

**Fig. 7 Creating a boundary line using a support vector machine**

● and × represent data points of two different groups depicted based on values of variables X1 and X2, respectively (e.g., laboratory test results). The line represents the boundary that maximizes the sum of the distances from the line.

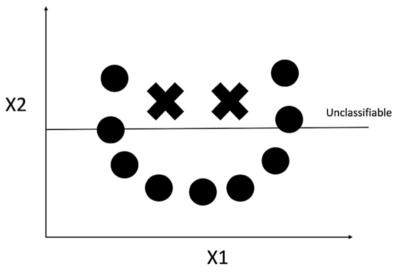

**Fig. 8 Cases where support vector machine cannot create a linear boundary**

● and × represent data points of two different groups depicted based on values of variables X1 and X2, respectively. The linear line cannot separate these data points.

trick can be used to map the data to a higher dimension called feature space [15, 17, 19] (**Fig. 9**). We can then separate these data points linearly. When the term SVM is used in medical literature, it usually refers to SVM with kernel trick (kernel SVM). Kernel SVM can deal with numerous variables, and it is easy to obtain good results, even with small data. On the other hand, it can be computationally expensive to process a large amount of data as the kernel trick generally increases the dimensionality [15, 17].

*Perceptron and neural network*

The perceptron is an algorithm that attempts to reproduce human-like cognitive abilities by imitating human neurons [20–22]. As shown in **Fig. 10**, perceptron adds a weight $w_i$ to the input data, passes it to the next stage (node or neuron), and then passes the obtained value to a function called an activation function, which outputs a predictive value when the sum of given values exceed a certain threshold. The weights are updated through the learning process: the output value is compared to the actual outcome (i.e., the right answer) and updated until the error becomes minimal.

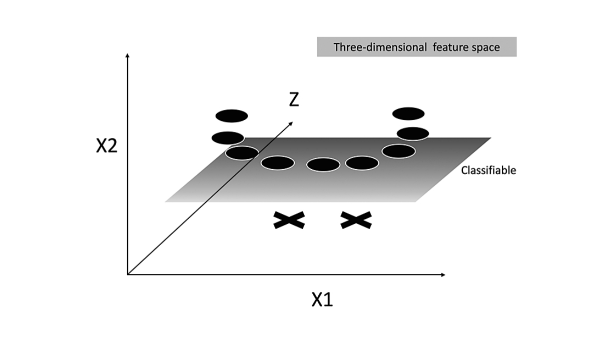

**Fig. 9 Boundary creation by kernel support vector machine**

● and × represent data points of two different groups depicted based on values of variables X1 and X2, respectively. The data are mapped in feature space for linear separation.

A neural network is a multi-layered combination of perceptrons, namely, the input layer, multiple hidden layers, and the output layer [21, 22] (**Fig. 11**). The number of hidden layers is a hyperparameter, a value that should be prespecified by a researcher before running the algorithm. By combining multiple layers, we can model more complex relationships than a simple perceptron. In the neural network, the most commonly used weight updating method is backpropagation, where the total error obtained from the current output is passed backwards and distributed to the preceding nodes in the hidden layers and then the ones in the input layer. The weights are adjusted by repeating this process to minimize the total error.

Deep learning typically refers to an advanced type of neural network that has multiple layers organized in deeply nested network architecture [23–25]. With using advanced operation, such as convolution for a digital image, and multiple activation functions in one node; deep learning achieves much better performance than a simple neural network. Deep learning is widely used in almost all the fields that embrace machine learning technology (e.g., medical imaging, natural language processing, speech and audio processing, and drug discovery). One of the deep learning methods that are particularly successful in medicine is a convolutional neural network for processing medical imaging [24]. Convolutional neural network makes a prediction based on the input of arrays of pixel intensities in the three-color channels.

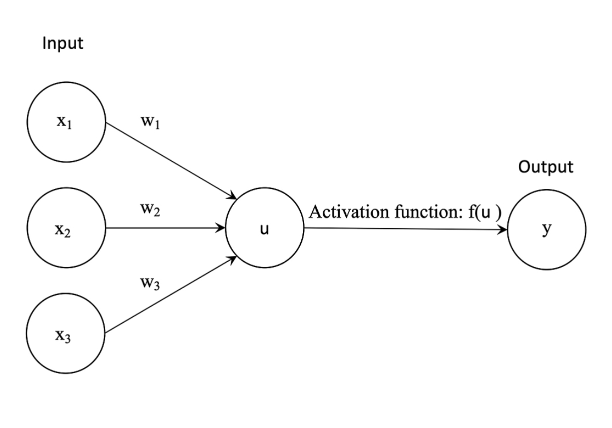

**Fig. 10 Perceptron**

$X_n$ represents variables (predictors) and $W_i$ represents weights.

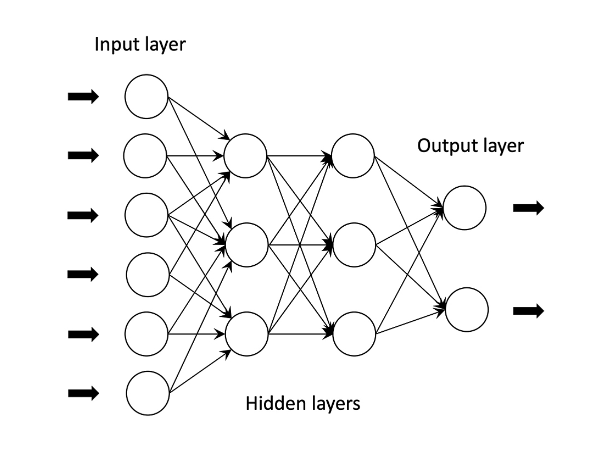

**Fig. 11 Neural network**

Diabetic retinopathy, for example, was identified with 90.3% of sensitivity and 98.1% of specificity by using a convolutional neural network model developed from retinal fundus photographs [26]. The other medical fields, where image classification demonstrated promising performance, were dermatology [27, 28], radiology [29, 30], pathology [31, 32] and cardiology [33–35].

### 2.3 Developing a Prediction Model Using Machine Learning

Although the algorithms of machine learning are much more complex than conventional methods, the sequence of steps for creating a prediction model is similar. The steps are to: (i) determine the research question, (ii) obtain data, (iii) preprocess the data and split them into training data and test data, (iv) apply the algorithm to the training data to develop a model, and (v) evaluate the performance of the model on the test data. The steps from step three onward are described in another article in this journal [36]. Data preprocessing in the step three includes imputation of missing values, creation of dummy variables, and normalization/standardization of data. Application in the step four is the main part of machine learning, and appropriate algorithms should be selected from various machine learning methods for each problem setting. To ensure accuracy of the prediction, one strategy is to try multiple machine learning methods and select the one with the best performance or to combine multiple methods to utilize the advantages of different methods.

Before applying the algorithm to data for machine learning, researchers have to specify hyperparameters. Hyperparameters are values that are external to the model and cannot learn from the data. Examples of hyperparameters are the number of decision trees to construct or features to select in a random forest. These hyperparameters affect the accuracy, complexity, and efficiency of the model. Because the model performance varies greatly depending on the values of hyperparameters, parameter tuning is required to find good values. The method called grid search is often used for hyperparameter tuning by manually or automatically changing the values of the hyperparameters little by little. When evaluating hyperparameters, cross-validation method is commonly used to improve the fit to unknown data by using "validation" data divided from training data (**Fig. 12**). The validation data here is for hyperparameter tuning; it differs from the one used for internal validation in the prediction model described in another article in this journal [36]. In the machine learning context, the dataset used for the internal validation is often called test data; while, it is called validation data in an epidemiological context.

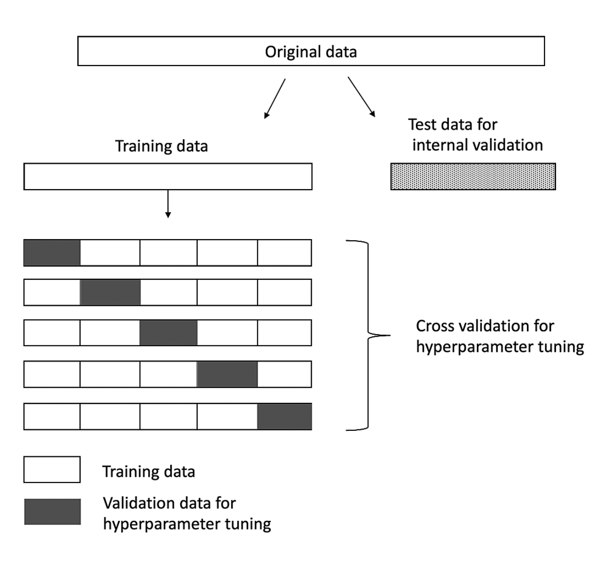
Original data
Training data
Test data for internal validation
Cross validation for hyperparameter tuning
Training data
Validation data for hyperparameter tuning

**Fig. 12** Cross-validation for hyperparameter tuning

In step five, the model developed in step four is applied to test data to evaluate its performance. As described in the article about the clinical prediction model [36], along with accuracy, sensitivity, specificity, and area under the curve are commonly used performance measures for classification problems. For regression problems that predict numerical value, root-mean-square error (RMSE) is used to evaluate the deviation of the predicted value from the observed (correct) values of the given dataset. If the prediction performance is unacceptably low, return to step two and reconsider each step.

## 3. SUPERVISED MACHINE LEARNING APPLICATIONS

### 3.1 <Example 1> Triage Systems Developed by Multiple Machine Learning Methods

Where medical resource is limited, differentiation and prioritization of critically ill children are important. Goto et al. [37] examined how well an objective triage system developed by machine learning can predict clinical outcomes of children presented to the emergency department (ED) compared to a conventional triage method based on a medical professional's assessment. The authors predicted in-hospital death and/or ICU admission using the least absolute shrinkage and selection operator or lasso in short, random forest, GBDT, and deep neural network. Variables used for prediction were age, sex, mode of arrival (walk-in vs ambulance), vital signs (temperature, pulse rate, systolic and diastolic blood pressure, respiratory rate, and oxygen saturation), visit reasons, patient's residence (home vs other [e.g., long-term care facility]), ED visits in the preceding 72 hours, and comorbidities. All the machine learning-based triage systems, although not statistically significant, performed better than the conventional triage system with a fewer number of undertriaged children. The authors concluded that machine learning-based triage systems may support clinicians in making triage decisions efficiently, thus improving optimal resource allocation.

### 3.2 <Example 2> Prognostic Prediction for COVID-19 Using a Combination of Machine Learning Methods

Health care systems worldwide are overwhelmed by the soaring number of COVID-19 patients. For early intervention and optimal resource allocation, an accurate prediction model for COVID-19 is needed. In contrast to Goto et al. [37] in example 1 where they evaluated multiple models separately, Gao et al. [38] integrated four different machine learning models into one ensemble model to predict the mortality risk of admitted patients for COVID-19. To develop the ensemble model, the authors first selected important 14 variables (consciousness, male, age, sputum, blood urea nitrogen, respiratory rate, D-dimer, number of comorbidities, platelet count, fever, albumin, $SpO_2$, lymphocyte, and chronic kidney disease) out of original 53 variables by lasso, another machine learning method often used for unimportant variable elimination. With the 14 variables, the authors developed 6 machine learning prediction models (logistic regression, SVM, GBDT, neural network, k-nearest neighbor, and random forests), then integrated the top 4 predictive models (logistic regression, SVM, GBDT, and neural network) into one. The ensemble model achieved an area under the curve of 0.96 and 0.92 for predicting mortality of COVID-19 in two external cohorts. The authors concluded that the model efficiently enables accurate risk stratification of COVID-19 patients on admission.

### 3.3 <Example 3> Classification of HIV Rapid Test Using Deep Learning

A rapid diagnostic test is a convenient and affordable option to screen for HIV in low- and middle-income countries. However, tests with weak or faint lines make visual interpretation diverse among field workers with different training levels. The accuracy of interpretation varied between 80% and 97%. Turbé et al. [39] developed a machine learning model to determine whether the results indicated positive or negative from photos of rapid diagnostic tests. A total of 11,374 images taken with tablets were labeled by three rapid diagnostic test experts and then used as a training dataset. The authors developed 4 models using the dataset; one is an SVM and three are different convolutional neural network models. One of the convolutional neural network models was used because of its best performance in terms of sensitivity and specificity. As a pilot test, the performance of the model was compared with those of 5 end-users with varying levels of training (2 nurses and 3 newly trained community health workers). In the visual interpretation of rapid diagnostic testing, the end-users' agreement levels were from 61 to 100%. The machine learning model demonstrated better performance than end-users for the following 4 indicators: sensitivity (95.6% vs. 97.8%), specificity (89% vs. 100%), positive predictive value (88.7% vs. 100%), and negative predictive value (95.7% vs. 98%). The authors concluded that rapid diagnostic testing images captured by a mobile device could standardize the interpretation of test results, reduce interpretation errors, and provide a platform for workforce training.

## 4. CHALLENGES IN MACHINE LEARNING

Machine learning is not a one-size-fits-all solution. Although the term "machine learning" gives the impression that everything is done automatically, it has some challenges. As with conventional methods, machine learning requires a good research question, sufficient sample size and variables, appropriate data sampling, and algorithm selection for each problem setting. When these processes are done heuristically by experts with domain knowledge, the model can achieve good performance. Automated machine learning without a human in the loop, especially in the medical field, has a risk to model artifacts because medical data often contain uncertainty, noise, and missing data [40].

An example is the failure of Google's influenza forecasting algorithm called Google Flu Trends. In 2008, Google developed an algorithm to quickly detect influenza trends from the combination of Google search terms data and actual survey data [41]. For the first few years, it surprisingly well predicted the number of cases, or trends of influenza, two weeks earlier and more accurately than the Centers for Disease Control and Prevention [42, 43]. Earlier prediction means being able to take action sooner, and it may even prevent future influenza pandemics. The introduction of this new technology has raised expectations for improving public health. However, several years later, the number of influenza cases predicted by Google's forecasting algorithm deviated significantly from the actual number of cases [44, 45].

Although Google did not provide a clear explanation for the suboptimal estimation, researchers speculated that the algorithm might have over-learned irrelevant search terms. Later, several studies analyzed how Google Flu Trends could have avoided erroneous forecasting by manually adding another data source or by updating the algorithm constantly [44, 45]. As in this example, machine learning sometimes produces unintended and erroneous results. When the public health policy was affected by such erroneous algorithms, what would be at stake were the lives of people. Regular human checks on these algorithms are therefore essential.

The application of machine learning to medicine raises another concern: the algorithmic predictions could control the "right" answer in the real world. For example, some physicians might blindly follow the prediction to admit patients who are classified as "inpatients" by the algorithm. While this behavior would further improve the apparent accuracy of the algorithm, it would obscure the true performance of the predictive algorithm. Furthermore, the physician, who is supposed to be the "teacher" in "supervised learning", would lose their credibility and authority if they are completely dependent on the "machine". The above concerns were expressed by several experts [40, 46], and there is still ongoing discussion on how to incorporate machine learning into clinical practice.

## 5. CONCLUSION

Machine learning has developed rapidly in the last decade with the improvement of computer performance and the advancement of statistics. This approach has massive potential for new insights into medicine given large numbers of variables, complex interactions, and nonlinear relationships between variables and outcomes. However, machine learning application in the medical field has only just begun. Owing to the complexity of medical domains, machine learning cannot fully substitute for human ability, at least for now. The combination of algorithms and human cognitive ability may be a key to the successful application of machine learning in medicine.

### CONFLICTS OF INTEREST

The authors declare no conflicts of interest in relation to the work presented in the manuscript.

### ACKNOWLEDGMENT

We would like to thank Masao Iwagami in the Department of Health Services Research, the University of Tsukuba, for his input and critical reading of the manuscript and feedback.

### REFERENCES

1. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* 2020;26:29–38.
2. Tokodi M, Schwertner WR, Kovács A, Tősér Z, Staub L, Sárkány A, et al. Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. *Eur Heart J* 2020;41:1747–56.
3. Wu W, Bleecker E, Moore W, Busse WW, Castro M, Chung KF, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. *J Allergy Clin Immunol* 2014;133:1280–8.
4. Lee S, Lee YH. Improving Emergency Department Efficiency by Patient Scheduling Using Deep Reinforcement Learning. Healthcare (Basel) [Internet]. 2020;8. Available from: http://dx.doi.org/10.3390/healthcare 8020077
5. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.
6. Kuhn M. Predictive Modeling with R and the caret Package. *User Model User-adapt Interact* 2013.
7. Kuhn M. The caret Package [Internet]. 2019 [cited May 10, 2022]. Available from: https://topepo.github.io/caret/
8. scikit-learn [Internet]. [cited May 10, 2022]. Available from: https://scikit-learn.org/stable/
9. Suthaharan S. Decision Tree Learning. In: Suthaharan S, editor. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. Boston, MA: Springer US; 2016: 237–69.
10. Tangirala S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *Int J Adv Comput Sci Appl* 2020;11:612–9.
11. Kotsiantis SB. Decision trees: a recent overview. *Artif Intell Rev* 2013;39:261–83.
12. Ying X. An overview of overfitting and its solutions. *J Phys Conf Ser* 2019;1168:022022.
13. Boulesteix A-L, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov* 2012;2:493–507.
14. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot* 2013;7:21.
15. Pisner DA, Schnyer DM. Chapter 6—Support vector machine. In: Mechelli A, Vieira S, editors. Machine Learning. Academic Press; 2020: 101–21.
16. Gunn SR, Others.. Support vector machines for classification and regression. *ISIS technical report* 1998;14:5–16.
17. Suthaharan S. Support Vector Machine. In: Suthaharan S, editor. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. Boston, MA: Springer US; 2016: 207–35.
18. Noble WS. What is a support vector machine? *Nat Biotechnol* 2006;24:1565–7.
19. Scholkopf B, Mika S, Burges CJC, Knirsch P, Muller K-R, Ratsch G, et al. Input space versus feature space in kernel-based methods. *IEEE Trans Neural Netw* 1999;10:1000–17.
20. Guyon I. Neural networks and applications

tutorial. *Phys Rep* 1991;207:215–59.

21. Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 2000; 43:3–31.

22. Islam M, Chen G, Jin S. An overview of neural network. *Am J Neural Netw Appl* 2019;5:7.

23. Deng L, Yu D. Deep Learning: Methods and Applications. Found Trends Signal Process. Hanover, MA, USA: Now Publishers Inc.; 2014;7:197–387.

24. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.

25. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electronic Markets* 2021;31:685–95.

26. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.

27. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.

28. Cho SI, Sun S, Mun J-H, Kim C, Kim SY, Cho S, et al. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network. *Br J Dermatol* 2020;182:1388–94.

29. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *Lancet Respir Med* 2018;6:837–45.

30. Wang G, Liu X, Shen J, Wang C, Li Z, Ye L, et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nat Biomed Eng* 2021;5:509–21.

31. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020;21:233–41.

32. Foersch S, Eckstein M, Wagner D-C, Gach F, Woerl A-C, Geiger J, et al. Deep learning for diagnosis and survival prediction in soft tissue sarcoma. *Ann Oncol* 2021; 32:1178–87.

33. Khurshid S, Friedman S, Reeder C, Di Achille P, Diamant N, Singh P, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation* 2022;145:122–33.

34. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019;394:861–7.

35. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related stroke. *Circulation* 2021;143:1287–98.

36. Iwagami M, Matsui H. Introduction to clinical prediction model. *Ann Clin Epidemiol* 2022;in press.

37. Goto T, Camargo CA Jr, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw Open* 2019;2:e186937.

38. 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;11:5033.

39. Turbé V, Herbst C, Mngomezulu T, Meshkinfamfard S, Dlamini N, Mhlongo T, et al. Deep learning of HIV field-based rapid tests. *Nat Med* 2021;27:1165–70.

40. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 2016; 3:119–31.

41. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* 2009;457:1012–4.

42. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting with Google Flu Trends. *PLoS One* 2013; 8:e56176.

43. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS One* 2011;6:e23610.

44. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science* 2014;343:1203–5.

45. Kandula S, Shaman J. Reappraising the utility of Google Flu Trends. *PLoS Comput Biol* 2019;15:e1007258.

46. Scott IA. Machine learning and evidence-based medicine. *Ann Intern Med* 2018; 169:44–6.