Family Practice, 2023, **40**, 200–204
https://doi.org/10.1093/fampra/cmac104
Advance access publication 1 October 2022
CASFM Methods Briefs

An introduction to machine learning for classification and prediction

Jason E. Black[1,2], Jacqueline K. Kueper[3,4], Tyler S. Williamson[1,2,5,*]

[1]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[2]O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[3]Department of Epidemiology and Biostatistics, Western University Schulich School of Medicine & Dentistry, London, ON, Canada
[4]Department of Computer Science, Western University Faculty of Science, London, ON, Canada
[5]Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

*Corresponding author: Department of Community Health Sciences, Centre for Health Informatics, Cumming School of Medicine, University of Calgary, 3280 Hospital Drive NW, Calgary, AB T2N 4Z6, Canada. E-mail: tyler.williamson@ucalgary.ca

Classification and prediction tasks are common in health research. With the increasing availability of vast health data repositories (e.g. electronic medical record databases) and advances in computing power, traditional statistical approaches are being augmented or replaced with machine learning (ML) approaches to classify and predict health outcomes. ML describes the automated process of identifying ("learning") patterns in data to perform tasks. Developing an ML model includes selecting between many ML models (e.g. decision trees, support vector machines, neural networks); model specifications such as hyperparameter tuning; and evaluation of model performance. This process is conducted repeatedly to find the model and corresponding specifications that optimize some measure of model performance. ML models can make more accurate classifications and predictions than their statistical counterparts and confer greater flexibility when modelling unstructured data or interactions between covariates; however, many ML models require larger sample sizes to achieve good classification or predictive performance and have been criticized as "black box" for their poor transparency and interpretability. ML holds potential in family medicine for risk profiling of patients' disease risk and clinical decision support to present additional information at times of uncertainty or high demand. In the future, ML approaches are positioned to become commonplace in family medicine. As such, it is important to understand the objectives that can be addressed using ML approaches and the associated techniques and limitations. This article provides a brief introduction into the use of ML approaches for classification and prediction tasks in family medicine.

Key words: algorithms, decision support systems, clinical, diagnosis, computer-assisted/methods, family practice, humans, machine learning

Many methods strive to estimate unknown current and future characteristics about an individual, such as simple heuristics (e.g. *anecdotal decision rules*) that have some ability to classify (categorical outcome) or predict (continuous outcome).[1] Statistical approaches such as regression modelling aim to better classify and predict using associations observed in data.[2] Similarly, machine learning (ML) provides a collection of tools to classify and predict based on patterns observed in data. Recently, ML has gained attention for classification and prediction tasks in health research, including family medicine research, due to the increasing availability of vast health data repositories (e.g. electronic medical record databases) and advances in computing power.[3–5] However, uncertainty and hesitation exist among clinicians and health researchers around the use of ML for health research objectives,[6] often due to insufficient ML expertise or concerns around model explainability, overfitting, and equity. This article aims to address these concerns by providing a brief introduction to ML for classification and prediction, with applications in family medicine.

What is ML?

ML describes the automated process of identifying ("learning") patterns in data to perform tasks, such as classification and prediction.[7] ML is a subfield of artificial intelligence, which considers how computers might "think" or process information "intelligently." Similar to familiar regression-based techniques, ML requires several user decisions, including specifying the outcome of interest for classification or prediction; selecting data to be used for learning the patterns; and determining the variables used to classify or predict.

ML approaches to classify and predict are commonly referred to as *supervised learning* techniques.[8] Other ML subfields include *unsupervised learning* and *reinforcement learning* and are described elsewhere.[2,9] Supervised learning uses data with known outcomes to learn how to classify or predict unobserved outcomes in new data. Often this is achieved by learning the associations between predictive variables and known outcomes, which are used to classify or predict outcomes among new individuals. For example, primary care electronic medical record data were used to predict premature death among people with epilepsy using an ML model[10]; the resulting model aims to identify individuals with epilepsy at high risk of premature death, allowing for earlier preventive interventions. While the current uptake of ML in family medicine has been limited,[11] important changes in family medicine are anticipated as ML approaches are introduced.[12]

Key messages

- Machine learning can classify and predict health outcomes.
- Potential exists in family medicine for risk profiling and decision support.
- This is a brief introduction to using machine learning in family medicine.

Common examples of supervised ML approaches are decision trees and random forests.[13] Decision trees repeatedly split data according to some predictive characteristics into 2 or more groups with increasingly homogenous outcomes. The data are first split into 2 or more groups according to the variable that maximizes the homogeneity (i.e. within-group similarity) of the resulting groups; subsequent splits are similarly performed within each resulting group until some stopping criterion is met, such as a threshold of homogeneity of outcomes or a prespecified maximum number of splits. The outcome of a new individual is predicted by evaluating the established splitting rules and assigning the most frequent outcome observed in the final group from the training data. Figure 1 demonstrates a decision tree for classifying risk of influenza developed by Afonso et al.[14] Extending from decision trees, random forests construct many trees while varying the observations and predictors included during the development of each tree. The outcome is classified or predicted based on the results of all trees. Other ML approaches used for classification and prediction tasks are described in Table 1.

How does ML differ from statistics?

Supervised ML originates from computer science, where it classifies or predicts outcomes, often based on learned associations with predictive variables. Beyond predicting and classifying, statistics are also commonly used to understand associations between potential causes and effects, particularly in the context of health research. ML and statistics are intimately related. Statistical techniques are sometimes augmented by ML approaches to accomplish a single research objective and vice versa. For example, in an analysis that weights observations according to the inverse of their probability of receiving some treatment to balance groups in terms of confounding variables, treatment probabilities may be predicted using an ML approach that prioritizes predictive performance, such as decision trees.[19] Statistical models can then use the weights determined by these ML-predicted treatment probabilities.

Both ML and statistical models can be used to develop classification and prediction models by summarizing the associations between an outcome and its predictors using parameter estimates. However, many ML approaches extend beyond parametric models; for example, similarity-based methods such as *k*-nearest neighbours can predict an unknown outcome for some new individual based on similar individuals whose outcomes are known without making any assumptions about the parametric structure of the data.

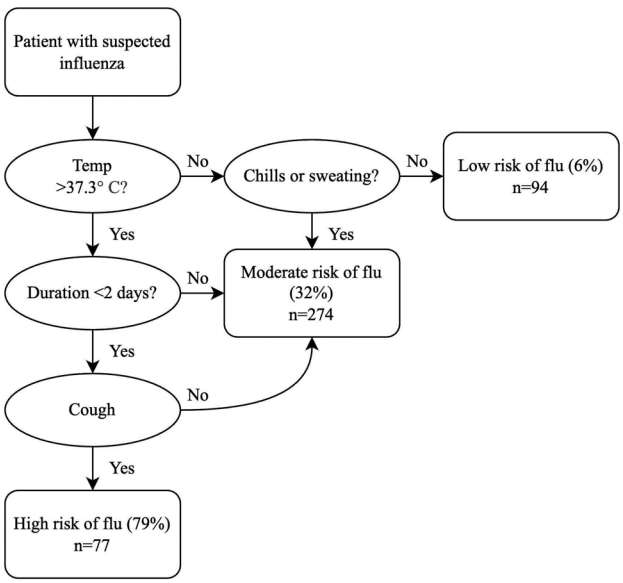

Fig. 1. Decision tree for influenza risk classification developed by Afonso et al. using 459 primary care or ambulatory care patients in Switzerland and the United States.[14]

Determining the best ML approach for classification and prediction

Multiple ML approaches may be suited to a given classification or prediction task. Typically, the best performing model may be desired, as identified by optimizing some measure of model fit or predictive performance.[20] Numerous measures can be optimized, such as accuracy, *F*1 score, area under the receiver operating characteristic curve, Akaike information criterion, Bayesian information criterion, coefficient of determination (R^2), root-mean-square error, and calibration.[2,7] Selection of which measures to use or prioritize will depend on several factors including user preferences and the intended use of the classification or prediction tool. For example, selecting the ML approach that fits the model with the greatest negative predictive value (NPV) maximizes the model's ability to correctly assign the label "negative"; this may be ideal for cancer screening tests where a false negative incorrectly indicates the patient is cancer-free and should be strongly avoided. In contrast, situations such as disease surveillance require balance between positive predictive value and NPV to provide accurate estimates of population disease burden.

Model training and selection should be performed on data that will not be used when testing the final model performance to avoid fitting random patterns in the data that may not be present in new data (i.e. overfitting).[21] One common approach splits data into 2 or more datasets: 1 dataset for model fitting or training; 1 for model selection; and 1 for model testing or validation, for example. Methods such as cross-validation and resampling procedures (e.g. bootstrapping) extend this approach and can be useful when fewer

Table 1. Common ML approaches for classification and prediction.

Approach	Description
Decision trees[13]	Data are repeated split into increasingly similar groups based on the variable that maximizes the similarity of the resulting groups ("splitting criterion"). New classifications and predictions are made by evaluating the new individual according to the established splitting criteria.
Random forests[13]	Many decision trees are constructed while varying the observations and variables included during tree development. New classifications and predictions are based on some consensus of the predictions from each component tree.
Support vector machines[15]	Data are divided into 2 or more groups based on an *n*-dimensional hyperplane that segments the feature space defined on *n* features. New individuals are assigned the outcome of the segment where they reside in the feature space.
k-Nearest neighbours[16]	New data are compared with data whose outcomes are known. The classified or predicted outcome of a new individual is determined based on the outcomes of the *k* individuals most similar to the new individual.
Neural networks[17]	A collection of connected nodes arranged in multiple layers are used to process information; each node receives a signal that is converted to an output. Weights are assigned to nodes to determine how they influence the final output.
Linear and logistic regression[18]	Relates 1 or more variables to some outcome based on associations observed in data. Linear regression finds the line that best fits the data according to some mathematical criterion. Logistic regression similarly finds the line that best fits the data after applying the logistic function.
Naive Bayes[18]	A simple probabilistic classifier that modifies an initial prior probability based on the observed frequency of features among cases and noncases in the data. All features are assumed to be independent (and hence "naive").
Ensembles[18]	Multiple ML models are combined to classify or predict. New classifications and predictions are based on some consensus of the predictions from each component model.

data are available to further reduce the risk of overfitting.[22] Temporal validation (i.e. assessing performance in the same population during a different time period) or external validation (i.e. assessing performance in a different population) are used to ensure the selected model performs strongly in new settings.[23]

Another factor that ought to be considered in selecting the ideal ML model are the properties of various approaches. For example, decision trees are intuitive and easy-to-use but prone to overfitting—in contrast with their extension: random forests. Random forests include many decision trees by varying the observations and variables included in model development and combines the results from each tree to predict the outcome better, at the cost of decreased transparency.[24] Such complex ML models often require information describing many predictors, incurring additional burden on users collecting such information, and require specialized tools (e.g. smartphone applications) that may not be practical for point-of-care use.

Additional specifications: hyperparameters

In regression, parameter estimates are determined by an automated process that optimizes some measure of model fit. For example, logistic regression determines the parameter estimates that maximize the log likelihood function. ML approaches may similarly learn parameters by optimizing an objective or loss function that describes how well the model fits the data, such as the log likelihood, hinge loss, or mean squared error. ML approaches frequently also require specification of 1 or more *hyperparameters* that determine some aspect of the model; the process of determining the optimal value for hyperparameter(s) is called "*tuning*."[25] Oftentimes models across a range of possible hyperparameters are fit to training data and the best one selected based on some measure of model fit or performance on validation data. For example, penalized regression models, such as the least absolute shrinkage and selection operator (lasso), require specification of a tuning parameter λ that corresponds to the amount parameters are reduced in magnitude to provide more conservative risk estimates (i.e. parameter shrinkage).[26] In penalized regression, cross-validation is commonly used to select the value of λ that optimizes model performance in new data by repeated splitting the data into 2 groups, training a model with some tuning parameter λ value, then evaluating model performance on the remaining group. The tuning parameter λ that has the best average performance across all validation folds is selected. Other examples of tuning parameters include the number and weighting of neighbours in *k*-nearest neighbours.

Strengths of ML approaches

ML approaches are primarily concerned with selecting the best performing model. As such, ML approaches can outperform the predictive performance of their statistical counterparts; however, this is not true in all applications. Any potential gains in model performance using ML approaches must be carefully weighed against differences in how easily the modelling process can be explained (i.e. model explainability) and general acceptance of the approach.

ML approaches are often more flexible than parametric statistical models. ML approaches can readily handle unstructured data (e.g. data with text information) by preprocessing the data into a structured form that can be included in the model.[27] Additionally, some ML approaches handle interactions between predictive variables rather adeptly compared with statistical approaches that can handle only a few prespecified interactions.[28]

Limitations of ML approaches

Some studies note near equivalent performance comparing ML and statistical approaches[29]—often traditional statistical approaches are sufficient for classification and prediction objectives and should be considered prior to ML approaches. Further, ML approaches typically require large datasets for model specification and estimation. Upwards of 200 events per candidate predictor may be required to estimate an ML model, compared with 20 events per candidate predictor using a statistical model.[30]

In some instances, ML approaches have been shown to be biased against specific—often marginalized—groups.[31] Indeed, errors in model design (e.g. label bias: label meanings are inconsistent across patient groups); biases in training data (e.g. missing data bias: data may be missing for marginalized groups); and issues when applying ML models (e.g. privilege bias: models may not be available where marginalized groups receive care) contribute to bias in ML applications.[32] ML approaches must be carefully applied to ensure equity in the resulting model and its applications.[33]

Certain ML approaches are criticized for their lack of transparency due to complexity that is inaccessible to most model developers and clinical users.[34] For example, neural networks process information repeatedly using several layers—where the output of 1 function is processed as the input for another—with many connections used to represent the relationships between predictors and some outcome. While these series of layers and connections may allow for highly accurate predictions, the model is not inherently interpretable so understanding the process to obtain predictions is nearly impossible. Explainable artificial intelligence is a broad field and area of ongoing research that includes developing inherently interpretable models, rendering approaches to better understand how "black box" ML models are behaving, and making ML models more accessible to nontechnical audiences.[35]

Lastly, calibration (i.e. the agreement between the estimated and observed number of events) and methods examining calibration of ML models have received little attention. Despite calls to evaluate calibration when developing classification and prediction models,[36] calibration is not prioritized among many ML approaches. This may have alarming impacts in family medicine, where treatment decisions may be based on classifications and predictions from ML models. Calibration must be evaluated to ensure accurate classifications and predictions, ideally in a new, external dataset.

Other ML applications

Extending beyond traditional classification and prediction, ML is frequently used to automate image classification. For example, ML models are trained to investigate images of skin abnormalities and flag those requiring further investigation.[37]

ML has also enabled the processing of large amounts of free text (e.g. unstructured clinical notes) using natural language processing (NLP). For example, NLP is used to process family physicians' clinical notes stored in electronic medical records to unearth more detailed patient information, including symptoms, lifestyle factors, and family history.[38]

Deep learning is a subfield of ML, where aspects of model development that would normally require manual specification are determined as part of the deep learning model.[39] Deep learning handles less structured data by automating how variables are processed by the model using multiple processing layers—lending the name "deep learning." These models are often not transparent or explainable.

Potential for ML in family medicine

In family medicine, classification and prediction objectives such as risk profiling and clinical decision support can be facilitated using ML approaches.[40] Risk profiling estimates a patient's risk of developing some disease or condition in the future based on models, including ML models, that use known patient information; high-risk patients can be targeted with risk-reducing interventions earlier in hopes of preventing or delaying disease onset. Clinical decision support can present additional information at times of uncertainty or high demand by providing appropriate suggestions or support that are learned using ML. For example, when deciding between treatment options, clinical decision support may help clinicians and patients by enabling the decision to be guided by tailored estimates of the probability of various outcomes.

Other opportunities for ML in family medicine beyond classification and prediction are anticipated. These include using ML to manage and synthesize information sources, such as scientific articles and clinical practice guidelines, to improve access to these vast resources for better, evidence-informed care. Clerical and routine tasks can be facilitated by ML, such as automated transcription of clinical notes, and can reduce administrative burden on clinicians to free their time for patient care.

Resources for those interested in implementing an ML approach are listed in Supplementary Table 1.

Summary

ML approaches hold promise to enable effective classification and prediction within family medicine. Compared with statistical approaches, using ML requires additional specifications around model selection and hyperparameter tuning and careful consideration of the strengths and limitations associated with each approach. In the future, ML approaches may become commonplace in family medicine; as such, it is important to understand the objectives that can be addressed using ML approaches and the associated techniques. This article provides a brief introduction into the use of ML approaches in family medicine research; however, further readings are required before pursuing an ML approach.

Acknowledgements

The authors thank Afonso et al. for their permission to reproduce their decision tree for influenza risk classification.

Supplementary material

Supplementary material is available at *Family Practice* online.

Funding

Jason Black is supported by the Achievers in Medical Sciences doctoral award, administered by the University of Calgary;

the Artificial Intelligence for Public Health trainee scholarship program; and Alberta Innovates.

Ethical approval

No ethics approval was required for this work.

Conflict of interest

None declared.

Data availability

No data were accessed for this work.

References

1. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science*. 1974;185(4157):1124–1131.
2. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. New York (NY): Springer; 2009.
3. Lin SA. Clinician's guide to artificial intelligence (AI): why and how primary care should lead the health care AI revolution. *J Am Board Fam Med*. 2022;35(1):175–184.
4. Liaw W, Kakadiaris IA. Primary care artificial intelligence: a branch hiding in plain sight. *Ann Fam Med*. 2020;18(3):194–195.
5. Pagliari C. Digital health and primary care: past, pandemic and prospects. *J Glob Health*. 2021;11:01005.
6. Bandyopadhyay A, Goldstein C. Clinical applications of artificial intelligence in sleep medicine: a sleep clinician's perspective. *Sleep Breath*. Published online March 9, 2022. doi:10.1007/s11325-022-02592-4
7. Mitchell TM. *Machine learning*. New York: McGraw-Hill; 1997.
8. Müller AC, Guido S. Introduction to machine learning with Python. O'Reilly Media, Inc.; 2016 [accessed 2022 May 4]. https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/
9. Sutton R, Barto A. *Reinforcement learning: an introduction*. 2nd ed. MIT Press; 2018 [accessed 2022 May 31]. http://incompleteideas.net/book/the-book-2nd.html
10. Hrabok M, Engbers JDT, Wiebe S, Sajobi TT, Subota A, Almohawes A, Federico P, Hanson A, Klein KM, Peedicail J, et al. Primary care electronic medical records can be used to predict risk and identify potentially modifiable factors for early and late death in adult onset epilepsy. *Epilepsia*. 2021;62(1):51–60.
11. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. *Ann Fam Med*. 2020;18(3):250–258.
12. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. *J Gen Intern Med*. 2019;34(8):1626–1630.
13. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81–106.
14. Afonso AM, Ebell MH, Gonzales R, Stein J, Genton B, Senn N. The use of classification and regression trees to predict the likelihood of seasonal influenza. *Fam Pract*. 2012;29(6):671–677.
15. Suthaharan S. Support vector machine. In: Suthaharan S, editor. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning. Integrated series in information systems*. New York: Springer US; 2016. p. 207–235.
16. Patrick EA, Fischer FP. A generalized *k*-nearest neighbor rule. *Inf Control*. 1970;16(2):128–152.
17. Bishop CM. Neural networks and their applications. *Rev Sci Instrum*. 1994;65(6):1803–1832.
18. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013 [accessed 2022 May 4]. https://link.springer.com/book/10.1007/978-1-4614-7138-7
19. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337–346.
20. Zucchini W. An introduction to model selection. *J Math Psychol*. 2000;44(1):41–61.
21. Dietterich T. Overfitting and undercomputing in machine learning. *ACM Comput Surv*. 1995;27(3):326–327.
22. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence—Vol. 2. IJCAI'95. Morgan Kaufmann Publishers Inc.; 1995. p. 1137–1143.
23. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu LM, Moons KG, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40.
24. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
25. Probst P, Boulesteix AL, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. *J Mach Learn Res*. 2019;20(53):1–32.
26. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267–288.
27. Guyon I, Elisseeff A. An introduction to feature extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA, editors. *Feature extraction: foundations and applications. Studies in fuzziness and soft computing*. New York: Springer; 2006. p. 1–25.
28. Boulesteix AL, Janitza S, Hapfelmeier A, Van Steen K, Strobl C. Letter to the Editor: On the term 'interaction' and related phrases in the literature on Random Forests. *Brief Bioinform*. 2015;16(2):338–345.
29. Nusinovici S, Tham YC, Yan MYC, Ting DS, Li J, Sabanayagam C, Wong TY, Cheng CY. Logistic regression was as good as machine learning for predicting major chronic diseases. *J Clin Epidemiol*. 2020;122:56–69.
30. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14:137.
31. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453.
32. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866–872.
33. Smith MJ, Axler R, Bean S, Rudzicz F, Shaw J. Four equity considerations for the use of artificial intelligence in public health. *Bull World Health Organ*. 2020;98(4):290–292.
34. Castelvecchi D. Can we open the black box of AI? *Nat News*. 2016;538(7623):20–23.
35. Holzinger A. From machine learning to explainable AI. In: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA). 2018. p. 55–66.
36. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
37. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, Gutman D, Halpern A, Helba B, Hofmann-Wellenhof R, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol*. 2019;20(7):938–947.
38. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, Shaw SY, Xia Z, Szolovits P, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*. 2015;350:h1885.
39. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
40. Kueper JK, Terry A, Bahniwal R, Meredith L, Beleno R, Brown JB, Dang J, Leger D, McKay S, Pinto A, et al. Connecting artificial intelligence and primary care challenges: findings from a multi stakeholder collaborative consultation. *BMJ Health Care Inform*. 2022;29(1):e100493.