Classification and predictive models using supervised machine learning: A conceptual review

M A Pienaar,[1,2] FCPaed (SA), Cert Crit Care (Paed, SA), PhD;
K D Naidoo,[3,4] MB BCh, DCH, FCPaed (SA), Cert Crit Care (Paed, SA), MMed (Paed)

[1] *Department of Paediatrics and Child Health, Division Critical Care, Faculty of Health Sciences, School of Clinical Medicine, University of the Free State, Bloemfontein, South Africa*
[2] *Paediatric Critical Care Unit, Universitas Academic Hospital, Bloemfontein, South Africa*
[3] *Division of Critical Care, School of Clinical Medicine, University of the Witwatersrand, Johannesburg, South Africa*
[4] *Intensive Care Unit, Chris Hani Baragwanath Academic Hospital, Johannesburg, South Africa*

***Corresponding author**: M A Pienaar (PienaarMA1@ufs.ac.za)*

Supervised machine learning models (SMLMs) are likely to be a prevalent approach in the literature on medical machine learning. These models have considerable potential to improve clinical decision-making through enhanced prediction and classification. In this review, we present an overview of SMLMs. We provide a discussion of the conceptual domains relevant to machine learning, model development, validation, and model explanation. This discussion is accompanied by clinical examples to illustrate key concepts.

South Afr J Crit Care 2025:41(1):e2937. https://doi.org/10.7196/SAJCC.2025.v411.2937

Contribution of the study
This conceptual review provides an overview and guide to the interpretation of SMLMs in the medical literature.

The treatment of critically ill patients involves numerous classification and prediction tasks. Consider a patient with acute appendicitis, a mean arterial blood pressure of 60 mmHg and a serum lactate of 3.1 mmol/L. The patient is **classified** (diagnosed) as having septic shock, based on the signs of infection and the presence of cardiovascular dysfunction, and treated accordingly.[1] Furthermore, it is entirely reasonable for the treating physician to anticipate the possibility of the patient subsequently developing acute respiratory distress syndrome (ARDS) or indeed dying. Given the negative impact of these possible outcomes, clinicians attempt to **predict** (estimate) the likelihood of these outcomes occurring, so that appropriate management can be put in place. These predictions are dependent on **learned** relationships between phenomena. That is, through training, experience and research, clinicians have **learned** that certain features reliably correlate with certain pathological states and outcomes.

Enhanced classification and prediction can decisively inform the delicate balance between benefit and risk inherent to the treatment strategies employed in the dynamic environment of the intensive care unit. If the risk of developing ARDS is substantial, then the early adoption of a lung protective ventilation strategy would be prudent despite the likely increased need for sedation.[2,3] Classification (stratification) of the severity of lung injury would likely inform the choice of oxygen target for the patient, while advances in genomic profiling promise a more predictable beneficial response to the use of corticosteroids, simultaneously avoiding unnecessary adverse effects in likely 'non-responders'.[4,5]

Conceivably, bedside physicians would perform these tasks deliberately, reasoning actively and giving these decisions their full attention. This considered process would predictably take some time, a resource in short supply when managing multiple tasks in parallel, within an environment often teetering on the edge of chaos. Realistically, we know that decisions are often made automatically, instinctively and 'on-the-fly'. While decisions taken by experienced clinicians are often correct, we must acknowledge the risks inherent to this current paradigm. This is referred to as System I (instinctive, fast or intuitive) thinking by Kahneman.[6] Thus, it is clear that clinical benefit would result from the development of approaches that can support classifications or predictions with System II characteristics (slow, careful thinking), but executed at speed, in the absence of an increase in the cognitive load on clinicians.

Artificial intelligence (AI) and machine learning (ML) offer the potential to support prediction and classification in practice. Currently, this is preferentially achieved with supervised machine learning.[7] While this is not the only role that AI and ML may fulfil, this is likely to be a prevalent approach in the literature and forms the focus of this review.

What is artificial intelligence and machine learning?

AI, in the simplest terms, refers to any intelligence in computer systems or machines. The study of this field spans multiple scientific and philosophical domains.[8-10] In relation to AI, ML is generally considered a subfield of AI that includes the algorithms developed in AI and uses these algorithms to perform specific tasks such as classification or prediction.[11]

This approach can be seen in relation to activities using data to understand phenomena or achieve certain goals, which is collectively

referred to as data science.[12] This scope also encompasses two other related fields – statistics and data mining.[13] There is some overlap between statistics, data mining and ML and, when used in concert, they are often able to reveal useful insights from data too vast, messy or complex for interpretation by the human mind alone. As such, it is useful in medical research to view these three fields as a cluster of techniques that can be employed to achieve certain tasks. Deep learning is a unique subset of ML that uses artificial neural networks (ANNs) to extract progressively more abstract relationships from data. This is particularly relevant to image classification, such as the application of ML to the measurement of ejection fraction on echocardiographic video.[14] The interactions between these fields are illustrated in Supplementary Fig. 1.

In ML, **learning** algorithms respond to patterns in data. Supervised learning algorithms use labelled data (supervision) during **training** to learn the relationships that informed the assigned labels. Subsequently, the trained model can then be applied to unlabelled data to generate classifications or predictions, consistent with the provided labels (Supplementary Fig. 2). In unsupervised learning, the algorithm attempts to independently/autonomously learn (derive) new relationships from unlabelled data to generate novel classifications or insights. In re-inforcement learning, the algorithm (agent) is trained to optimise the achievement of a specific task using punishment and reward – akin to how one would approach playing a video game. While this approach is currently not widely represented in medical literature, it may see increased relevance in the automation of personalised haemodynamic or respiratory therapy.[15]

In ML, learning algorithms (Box 1) are sets of procedures that can iteratively alter the internal parameters of a mathematical model based on training data to optimise the accuracy of its subsequent predictions (generally by minimising error).

The widely used APACHE II[16] and PIM3[17] predictive scores are informed by logistic regression models, relatively simple examples of learning algorithms. In contrast, ANNs are more complex examples of algorithms (models) that achieve learning by loosely mimicking the neuron pools within the human brain (Supplementary Fig. 3). In the same way that neuron pools adjust the strength of impulses and activation thresholds to control propagation based on different stimuli, ANNs adjust the weights of connections between neurons to minimise error based on training data.[18] Regardless of the specific algorithm utilised (Box 1), algorithms aim to learn the relationships within data with little direct explicit programming.[18]

Building a supervised machine learning model (SMLM)

Once the goal of the machine learning model (MLM) has been determined, building a useful SMLM begins. This encompasses two distinct phases or processes: development and validation. Development

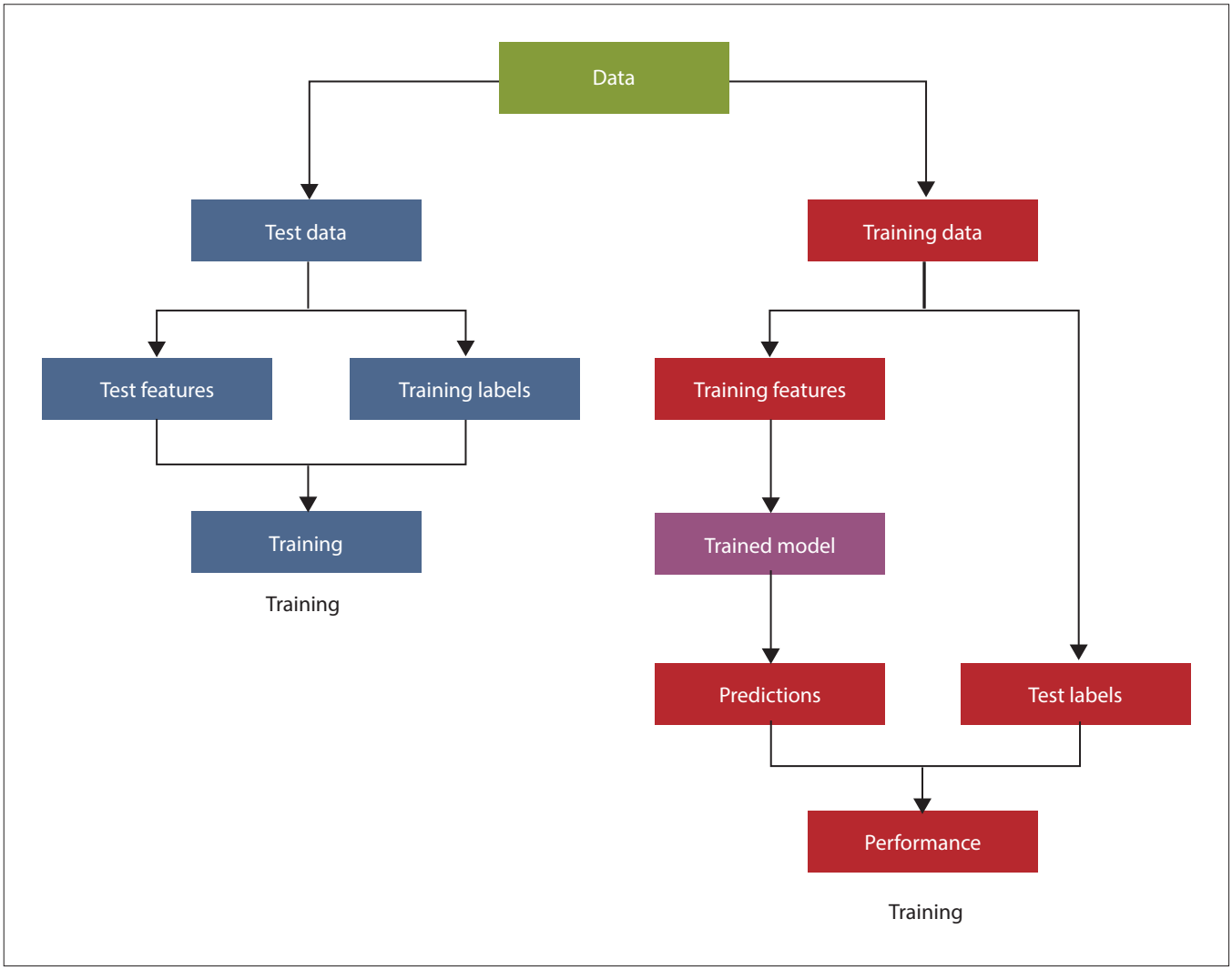

Fig. 1. Training and testing in supervised learning.

Box 1

Examples of machine learning algorithms
Linear regression
Logistic regression
Decision trees
Random forests
Gradient boosting and extreme gradient boosting
Artificial neural networks
Naïve Bayes

speaks to the training (learning) of the model, to enable the performance of the prediction or classification of interest. Validation (testing) subsequently ensures that the outputs of the model are acceptable, i.e. fit for purpose. Importantly, validation usually incorporates testing the model's performance with 'unseen' data, i.e. data that were not used during the development phase.

The first step is therefore to acquire the necessary data, which are divided into training and testing data sets respectively. By convention, the training set tends to be larger than the testing set, while remaining cognisant of the fact that both development and validation are optimised by the magnitude of the datasets afforded to each process respectively. Reasonable trade-offs between development and validation datasets are generally in the ratio of 3 - 4:1 (approximately 75 - 80% data allocated to development).[19]

Training and test data are then further divided into features (independent variables) and labels (dependent variables). Labels are provided by a human expert and are considered the ground truth (y). During training, the algorithm learns to minimise error in predicting the labels. This process yields the trained model which is now ready to make predictions. In testing, the goal is to use the features in the test data to make predictions ($\hat{y}$) using the trained model. By benchmarking these predicted labels against the provided test labels, we can estimate the predictive performance of the model (Fig. 1).

Feature selection

To be successfully trained, learning algorithms requires features that are associated with, and therefore predictive of, a particular label. While some features ultimately prove useful, others may be redundant. The inclusion of redundant features increases the complexity of models and may make them less generalisable and/or accurate.[20] Large sets of features may also increase the costs of acquiring data and limit their utility in practice. To that end, choosing appropriate features is a crucial contributor to the success of a model.

Various statistical and ML techniques can be used to evaluate which features are likely to be useful. In medical applications, clinical domain knowledge can also be integrated into feature selection and designing informative features that can predict the target feature.[21] For example, known features of organ dysfunction were important in the development of the sequential organ failure assessment (SOFA) score. Domain knowledge is likely the key factor underpinning the design of datasets or data collection tools which ensure the efficient collection of meaningful data.

Data preparation

Data need to be prepared for training. Preparation begins by inspecting the data in tables and through visualisation, to evaluate the distributions, outliers, missing data, relationships and correlations in the data. Outliers should be inspected to determine if they are related to errors or true values.

Missing data is a common problem which may influence training. This can be addressed by various approaches. The most intuitive would entail recovering the data from its source. Unfortunately, this is often a laborious undertaking, particularly when dealing with large datasets. The simplest alternative approaches include either deletion of records with selected missing data, or deletion of features with missing data. Note that the resultant loss of data may be a major impediment to training, especially where smaller datasets are being used. Techniques to impute missing data could employ either the mean, median or mode, or even prior or subsequent values. Multivariate approaches, such as K-nearest neighbours and multiple imputation by chained equations, represent examples of more advanced options used to impute missing data from other features through the analysis of the available data and features within the dataset.[22]

It is important to realise that MLMs require numerical inputs. Consequently, categorical features need to be encoded. Similarly, data which exist on different quantitative scales create challenges, as ML algorithms generally require values to be presented on a comparable scale. This is often achieved by scaling values to [-1 to 1] or [0 to 1].[23]

Hyperparameter tuning

While ML algorithms learn with minimal explicit programming, they do have parameters which require optimisation (tuning); these are called hyperparameters and are set before training. These features are specific to each algorithm. The approach to setting these parameters is to some extent determined by good practice, but often involves a grid search where an array of possible hyperparameters is tested in their possible permutations through cross-validation (see below). Following the tuning and selection of the best hyperparameters for the algorithm of interest, the MLM is then ready to be trained.[24]

Cross-validation

At this point, the model may be tested directly on the test data, but often, when the availability of data is limited, it is pragmatic to obtain a preliminary estimate of model performance before moving on to validation. This is relevant where testing data may be too small to provide a generalisable estimate of model performance. Cross-validation involves serially splitting the training data into different training-testing splits called folds (Fig. 1). In each fold, the model is trained and tested, and the performance metrics are averaged.[25] If this performance is satisfactory, the model is retrained on the whole training set and the validation study can commence.

Model validation

Broadly speaking, the two outputs of SMLMs are either the allocation (assignment) of an instance to a class (or classes), or the generation of a predicted probability. With respect to class assignment, the instance now has a value of 1 for the class/es that it belongs to and simultaneously a value of 0 for all classes it does not belong to. Models may also output a predicted probability, that is the extent to which the predicted class is likely to occur as a proportion of events with that predicted probability.[23]

By combining the basic outputs, SMLMs are also able to classify instances based on predicted probabilities by setting a threshold probability relative to which an instance will be classified into a group. Despite the widespread usage of the term 'critically ill', there is currently no consensus definition for the term.[26] This creates difficulty when

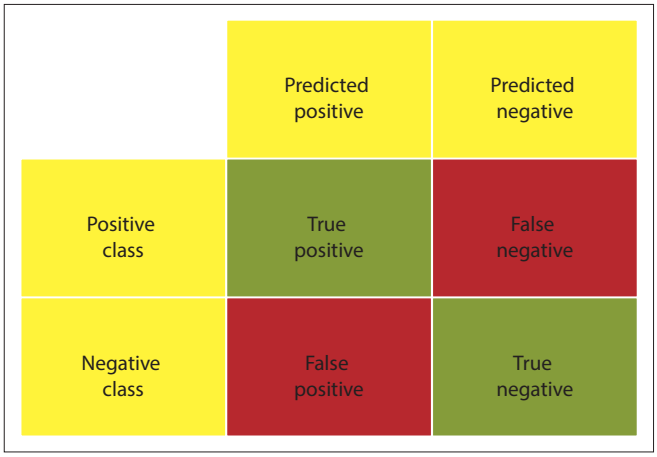

Fig. 2. Confusion matrix.

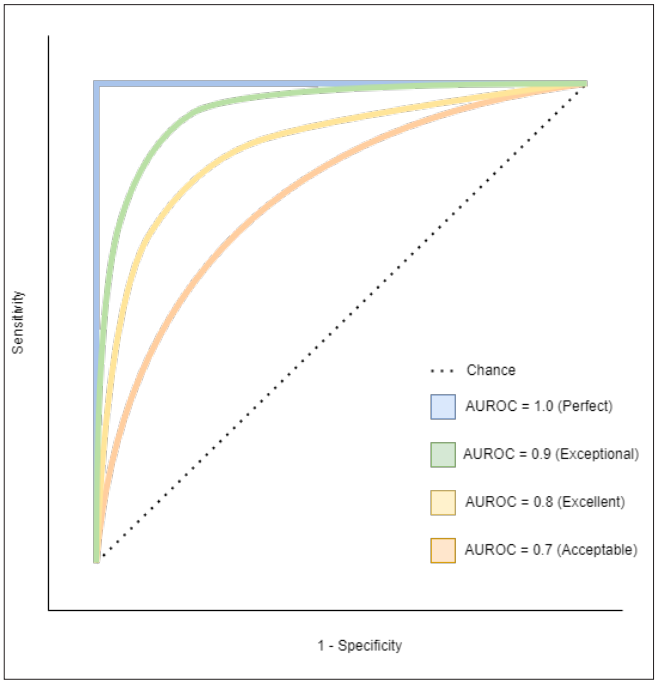

Fig. 3. Example ROC curves. (ROC = receiver operating characteristic curve; AUROC = area under the ROC.)

attempting to perform studies in 'critically ill' patients. Given that many consider 'critical illness' to represent the presence of a significant mortality risk, an SMLM trained to predict the mortality risk of patients would then be able to classify patients into the classes 'critically ill' or 'not critically ill' based on their predicted mortality risk compared with a pre-assigned threshold (e.g. 10%). The model would classify all patients with a mortality risk in excess of 10% as 'critically ill' and those with a risk of 10% or less as 'not critically ill.'

Validation is the process of quantifying the performance of a model to assess suitability for its intended use. Validation of the SMLM is thus focussed on the reliability of the classifications (discrimination) and the accuracy of the predicted probabilities (calibration).[27,28] Models exhibiting perfect performance are a utopic goal. Indeed, the pursuit of perfection during training comes at the inevitable cost of reduced generalisability (overfitting). Therefore, in reality, models are expected to make some errors.[27,28] Consequently, these errors need to be quantified to allow for the provision of realistic expectations to the end-user. Ultimately, for a model to be usable in practice, the performance of the model should be both known and satisfactory.

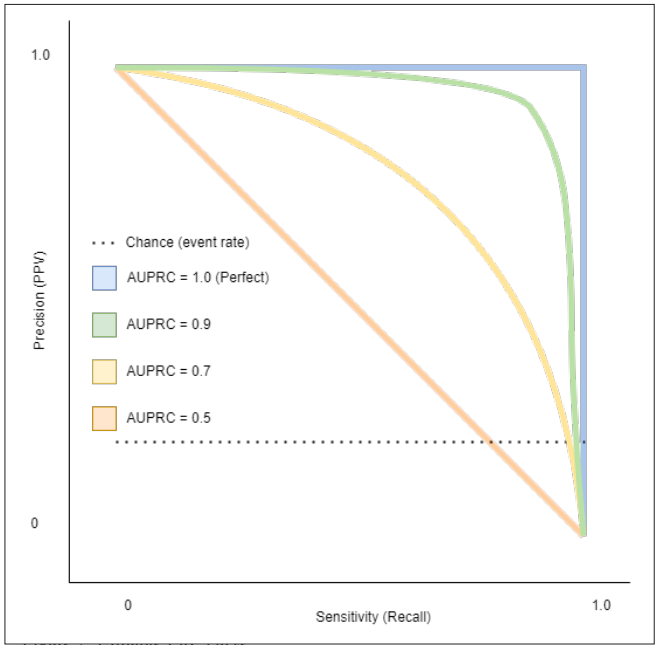

Fig. 4. Example PRC curve. (PRC = precision-recall curve; AUPRC = area under the PRC.)

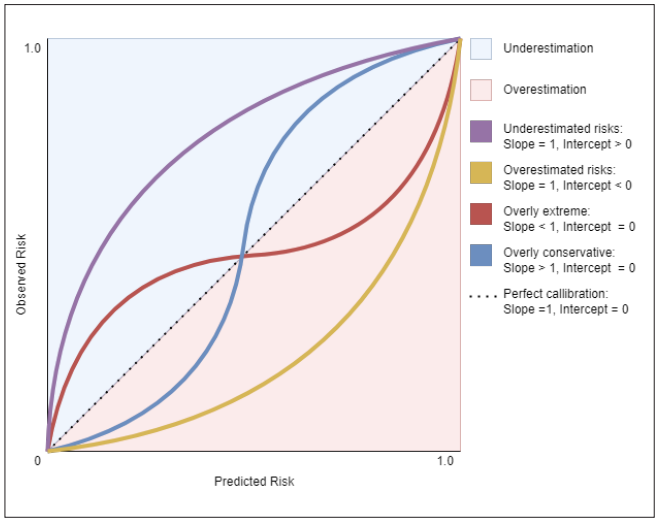

Fig. 5. Flexible calibration curves. In this figure, underestimated and overestimated risk and overly extreme models are contrasted with the perfect calibration (dotted) line.

Discrimination

Discrimination refers to the extent to which a prediction separates two classes of a dependent variable. If we imagine a perfect biomarker for bacterial infection, we expect that all patients who have a positive test will have bacterial infection, and all patients with a negative test will not. However, the existence of both false negatives and false positives is unavoidable. The relationships between the ground truth and the (mis)classifications produced by an associated test generates a special contingency table referred to as the confusion matrix (Fig. 2).

From the confusion matrix, we derive the performance metrics for classifications (Box 2). Clinicians will be familiar with the concepts of the contingency (2x2) table, sensitivity, specificity and positive predictive value from diagnostic tests.

Where probabilities are used to classify cases, this will lead to a range of probabilities between 0 and 1 (0 - 100%) which could serve as the threshold for classification. As this threshold is increased from 0,

the sensitivity would decrease from 1 to its minimum. At the same time, specificity and precision will increase. This allows estimation and comparison of model discrimination across the entire model by reflecting this compromise.

The most common model-wide estimate of discrimination is the receiver operating characteristic curve (ROC) which depicts the compromise between the sensitivity and specificity of model predictions (Fig. 3). The area under the ROC (AUROC) provides the probability that the predicted probability of an outcome will be higher in individuals with the outcome of interest than in those without. The AUROC value of a random classifier (coin toss) is 0.5. A minimum AUROC of 0.7 is considered acceptable. Values above 0.8 and 0.9 are generally considered excellent or exceptional respectively (28 - 30).

Frequently in medicine, most participants in a data set will not have the outcome of interest. For example, in children with sepsis, mortality is considerably less likely than survival.[31] Thus, the outcome of mortality will be the **minority class**. In situations where class imbalance exists, the ROC is likely to present an overoptimistic estimate of discrimination (suggesting greater performance than is truly present). In these cases, the precision-recall curve (PRC) may provide a better estimate of performance (Fig. 4).[32,33] The PRC plots precision against sensitivity, metrics which are both positively associated with the outcome of interest. This promotes improved analysis of minority class data. The area under the PRC (AUPRC) for a perfect model would be 1.0.[32–34] Thresholds for assessing AUPRC are not defined but can be considered in the context of the scenario, the event rate and by comparing AUPRC between models. A non-discriminating model would have an AUPRC less than or equal to the known event rate (usually represented by a horizontal line) on the PRC plot.[32] For example, if the rate of mortality in a population is 10%, a non-discriminating model would have an AUPRC of 0.1, while a model with an AUPRC of 0.65 would have discriminatory value and be considered superior to a model with an AUPRC of 0.5. Fig. 4 additionally illustrates that optimised performance tends toward the top right, in contrast to the ROC, which is the top left.

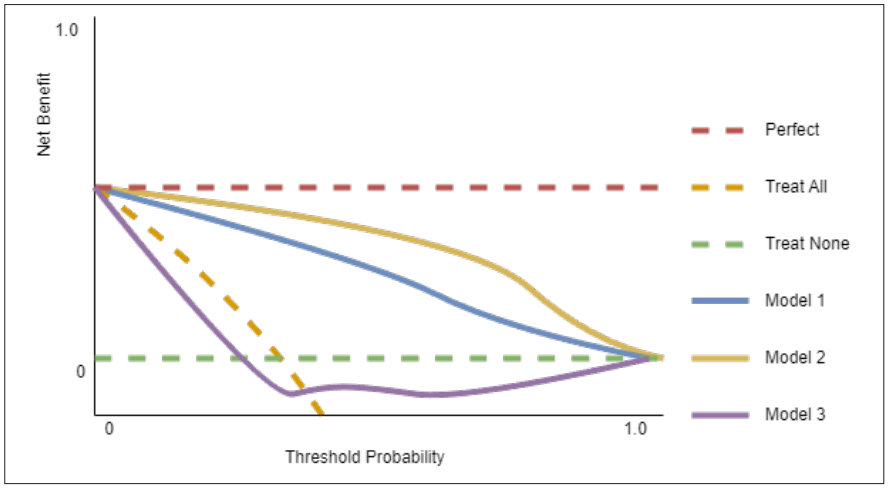

Fig. 6a. Decision curve analysis.

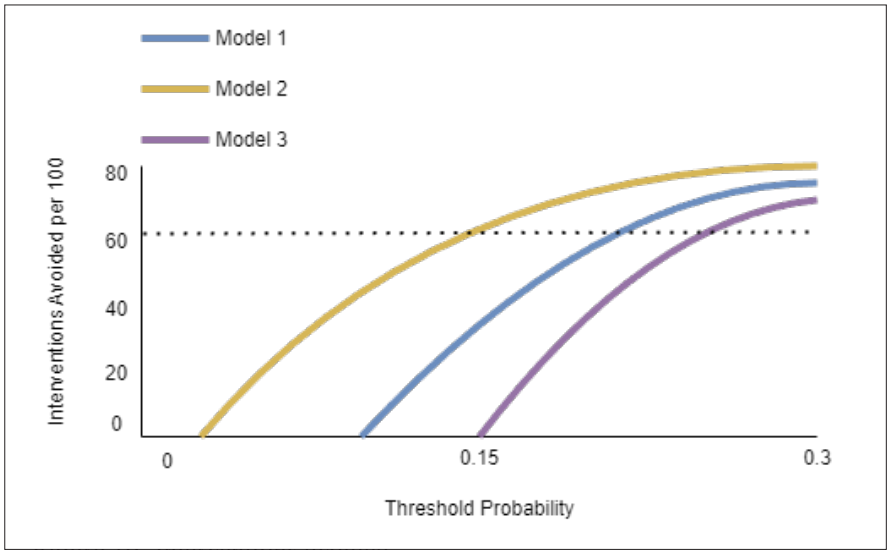

Fig. 6b. Interventions avoided.

Calibration

Calibration refers to the determination of how accurate the probabilities predicted by the SMLM are. This is accomplished by comparing the **predicted** probabilities of the outcome of interest to **measured (actual) probabilities** within the validation dataset. For example, consider the performance of two models developed to predict the need for renal replacement therapy (RRT) in patients with rhabdomyolysis using serum creatinine. Both models are applied to a validation data set containing 1 000 patients with rhabdomyolysis. Model A predicts that the probability of RRT

Box 2. Classification metrics

Metric	Calculation	Meaning
Sensitivity/recall/true positive rate	True positives/(true positives + false negatives) (a)/(a+b)	The probability of a positive test or prediction if the condition is present.
Specificity/true negative rate	True negatives/ (false positives + true negatives) (d)/(c+d)	The probability of a negative test or prediction if the condition is absent.
Precision/positive predictive value (PPV)	True positives/ (true positives + false positives) (a)/(a+c)	The probability of the condition is present in the case of a positive test or prediction.

in patients with creatinines 2 000 - 3 000 μmol/L is 0.1 and is 0.3 when creatinine exceeds 10 000 μmol/L. In comparison, Model B predicts probabilities of 0.2 and 0.4, respectively. Within the validation dataset, there are 240 patients with creatinines of 2 000 - 3 000 mol/L (23 required RRT) and 100 patients with creatinines greater than 10 000 μmol/L (30 required RRT). With respect to these two ranges of serum creatinine, Model A predicted RRT need in 24 and 30 patients, while Model B predicted 48 and 40, respectively. When compared with the actual values, Model A appears better calibrated than Model B. As such, calibration is the degree to which predicted probabilities agree with real probabilities for an outcome.[28]

Calibration can be assessed by examining the mean predicted probabilities of a model, the event rate, and the flexible calibration curve, which represents observed probability (prevalence) of a condition on the y-axis at each predicted probability on the x-axis.[27,35,36] Depending on the size of the data, this is achieved by breaking up the data into bands of predicted probability called bins. Calibration can be, at the lowest level, the agreement between the mean probability of a model and the event rate. In weakly calibrated models, the slope of the flexible calibration curve (Fig. 5) is close to 1 and the intercept is close to 0. In moderately calibrated models, the prior condition is met, and the calibration curve is close to the perfect calibration line from [0,0] to [1,1]. Strong calibration refers to the idealistic goal of near-perfect calibration of predictions to event rates in all categories of prediction.[36] Simplified examples of flexible calibration curves are provided in Fig. 5. This approach provides a more robust assessment of calibration over a range of probabilities than the widely reported Hosmer-Lemeshow statistic.[27]

Decision curve analysis

In the current paradigm pertaining to the use of corticosteroids in ARDS, two strategies are available, i.e. provide all ARDS patients with corticosteroids (treat-all) or provide corticosteroids to no-one (treat-none). In the treat-all population, the steroid responders enjoy the benefit of the treatment and experience the adverse effects, with the expectation that the benefit outweighs the harm. Conversely, the non-responders will only experience the adverse effects of the corticosteroids without any potential for benefit. On the other hand, a treat-none approach would protect all patients from the adverse effects but still causes harm by virtue of denying the responders the benefit of the intervention.

Decision curve analysis (DCA) is an approach designed to estimate the net benefit of a predictive model over a range of threshold probabilities and simultaneously compare its performance against the two opposing baseline strategies of treat-all v. treat-none. This approach combines aspects of both discrimination and calibration by employing the following equation:

$$Net\ benefit = sensitivity \times prevalence - (1 - specificity) \times (1 - prevalence) \times w$$

where w is the odds at the threshold probability. The higher the net benefit, the greater the performance of the model.

Consider the DCA of three hypothetical models assessing the use of corticosteroids in ARDS. When comparing models using DCA, the curves of better models are higher on the y-axis and further to the right on the x-axis. Fig. 6A illustrates that model 3 performs poorly across the range of threshold probabilities whereas both models 1 and 2 demonstrate better performance than either the treat-all or treat-none strategies. Overall, model 2 provides the greater net benefit.

The second notable benefit provided by DCA is the ability to determine the number of interventions avoided across the range of thresholds. This quantifies the number of likely non-responders who will be spared the adverse effects of the intervention. From Fig. 6B, at a threshold probability of 0.15 (to provide corticosteroids), model 2 avoids more than 60% of unnecessary exposure to steroids, without missing patients likely to benefit. At the same threshold, model 1 limits exposure in 40% of patients, while model 3 is unable to confer any benefit in this regard. Thus, model 2 has demonstrated the best performance in terms of maximising the potential benefits of corticosteroid administration within an ARDS population, while minimising unnecessary exposure (iatrogenic harm). The reader is directed to recent reviews for a more comprehensive description of DCA methodology.[37,38]

Model explanation

Clinicians need to understand why a model makes a specific prediction. This understanding underpins the trust in systems or may be relevant in audit chains, fault reporting, or circumstances where experienced clinicians may override recommendations. This is the role of the clinician in deployment. Holzinger refers to this as the 'doctor-in-the-loop'.[39] In our hypothetical corticosteroid example, the model may indicate that it is predicting a higher likelihood of response to corticosteroids **because** of the presence of markers of inflammation or a high oxygenation index.

Some models are more transparent, e.g. linear regression, logistic regression and tree-based algorithms provide explicit explanations for how features relate to predictions. These may be referred to as glass boxes and these explanations are referred to as *ante hoc* (beforehand).[11] Other algorithms are less transparent in their workings, particularly in the case of ANNs. Transparency, however, often comes at the cost of performance.[40] While these less transparent (black-box) models cannot provide *ante-hoc* explanations, these explanations can be provided *post-hoc* through further analysis. For example, Shapley Additive Explanations is a method to measure or quantify the contribution of each feature to a model prediction and present them graphically.[41] This representation aids in understanding how models evaluate each feature and correlate them with clinical or causal knowledge of disease or outcomes. This approach can be used for any kind of algorithm (this can be referred to as being model agnostic).

Conclusion

Supervised machine learning is likely to have an important role in future clinical practice by improving diagnostic performance and predictive analytics. The success of clinical machine learning applications will stem from a foundation of sound methodology in model development, validation and explanation.

Declaration. None.

Acknowledgements. The authors would like to extend their acknowledgement to Prof. Stephen Brown (University of the Free State), Dr Elizebeth George (University College London, Prof. Nicolaas Luwes (Central University of Technology) and Dr Jospeh Sempa, University of the Free State, who contributed to the authors' understanding of these concepts. We would also like to thank Dr Jacques Maritz (University of the Free State) for his inputs in the concepts of artificial intelligence and machine learning.

Author contributions. MAP contributed to conceptualisation, writing and preparation of the manuscript. KDN contributed to conceptualisation, critical appraisal and preparation of the manuscript.

Funding. None.

Conflict of interest. None.

1. Singer M, Deutschman CS, Seymour C, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315(8):801-810 (accessed 25 November 2024). https://jamanetwork.com/journals/jama/fullarticle/2492881
2. Qadir N, Sahetya S, Munshi L, et al. An update on management of adult patients with acute respiratory distress syndrome: An official Am Thoracic Soc Clin Pract Guide 2023;209(1):24-36. https://doi.org/101164/rccm202311-2011ST
3. Emeriaud G, López-Fernández YM, Iyer NP, et al. Executive summary of the Second International Guidelines for the Diagnosis and Management of Pediatric Acute Respiratory Distress Syn-drome (PALICC-2). Pediatric Crit Care Med 2023;24(2):143-168.
4. Buell KG, Spicer AB, Casey JD, et al. Individualised treatment effects of oxygen targets in mechanically ventilated critically ill adults. JAMA 2024;331(14):1195-1204 (accessed 25 November 2024). https://jamanetwork.com/journals/jama/fullarticle/2816677
5. The National Heart L and BIARDS (ARDS) CTN. Efficacy and safety of corticosteroids for persistent acute respiratory distress syndrome. N Engl J Med 2006;354(16):1671-1684 (accessed 25 November 2024). https://www.nejm.org/doi/full/10.1056/NEJMoa051693
6. Kahneman D. Thinking, fast and slow. Reflections on the liar. New York: Farrar, Straus and Giroux; 2011:499.
7. Ajibade SSM, Alhassan GN, Zaidi A, et al. Evolution of machine learning applications in medical and healthcare analytics research: A bibliometric analysis. Intelligent Systems with Applications 2024;24:200441.
8. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabol 2017;69:S36-S40.
9. Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 2000;44(1-2):207-219.
10. Turing AM. Computing machinery and intelligence.Mind 1950;49.
11. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):1-13.
12. Peng RD, Matsui E. The art of data science. Victoria, Canada: Leanpub Publishing; 2017.
13. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317-1318.
14. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580(7802):252 (accessed 25 November 2024). https://pmc.ncbi.nlm.nih.gov/articles/PMC8979576/
15. Mohri M, Afshin Rostamizadeh AT. Foundations of machine learning, 2nd ed. Vol. 60, Statistical Papers 2019.
16. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: A severity of disease classification system. Crit Care Med 1985;13(10):818-829.
17. Straney L, Clements A, Parslow RC, et al. Paediatric index of mortality 3: An updated model for predicting mortality in pediatric intensive care. Pediatric Crit Care Med 2013;14(7):673-681.
18. Graupe D. Principles of artificial Neural Networks. 3rd ed. Singapore: World Scientific Publishing; 2013.
19. Splitting a dataset into train and test sets | Baeldung on computer science (accessed 28 April 2022. https://www.baeldung.com/cs/train-test-datasets-ratio
20. Laksana E, Aczon M, Ho L, Carlin C, Ledbetter D, Wetzel R. The impact of extraneous features on the performance of recurrent neural network models in clinical tasks. J Biomed Inform 2020;102:103351. https://doi.org/10.1016/j.jbi.2019.103351
21. Pienaar MA, Sempa JB, Luwes N, George EC, Brown SC. Elicitation of domain knowledge for a machine learning model for paediatric critical illness in South Africa. Front Pediatr 2023;11:1005579.
22. Kotsiantis S, Kanellopoulos D, Pintelas PE. Data preprocessing for supervised learning. Int J Comp Inf Eng 2007;1(12):4104-4109. https://www.researchgate.net/publication/228084519
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. 2 January 2012. http://arxiv.org/abs/1201.0490
24. Claesen M, De Moor B. Hyperparameter search in machine learning. 2015 (accessed 4 April 2022). https://www.codalab.org/competitions/2321
25. 3.1. Cross-validation: Evaluating estimator performance — scikit-learn 1.1.1 documentation (accessed 1 June 2022). https://scikit-learn.org/stable/modules/cross_validation.html
26. Kayambankadzanja RK, Schell CO, Wärnberg MG, et al. Towards definitions of critical illness and critical care using concept analysis. BMJ Open 2022;12(9):e060972 (accessed 17 January 2025). https://pubmed.ncbi.nlm.nih.gov/36606666/
27. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. BMC Med 2019;17(1).
28. Schmid CH, Griffith JL. Multivariate classification rules: Calibration and discrimination. In: Encyclopedia of Biostatistics. Hoboken, USA: John Wiley & Sons; 2005.
29. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thoracic Oncol 2010;5(9):1315-1316.
30. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 2007;115(5):654-657.
31. Sanchez-Pinto LN, Bennett TD, Dewitt PE, et al. Development and validation of the Phoenix criteria for pediatric sepsis and septic shock. JAMA 2024;331(8):675-686 (accessed 16 January 2025). https://jamanetwork.com/journals/jama/fullarticle/2814296
32. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):1-21.
33. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. ACM International Conference Proceeding Series. 2006;148:233-234.
34. Jeni LA, Cohn JF, Torre FD LA. Facing imbalanced data--recommendations for the use of performance metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013. p. 245-251.
35. Van Calster B, Vickers AJ. Calibration of risk prediction models. Medical Decision Making 2015;35(2):162-169.
36. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: From utopia to empirical data. J Clin Epidemiol 2016;74:167-176.
37. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3(1):1-8 (accessed 12 March 2022). https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-019-0064-7
38. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making 2006;26(6):565-574 (accessed 12 March 2022). https://journals.sagepub.com/doi/10.1177/0272989X06295361
39. Holzinger A. Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inform 2016;3(2):119-131.
40. Holzinger A. From machine learning to explainable AI 2018. https://hci-kdd.org
41. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;December:4766-4775 (accessed 6 May 2022). https://arxiv.org/abs/1705.07874v2

Received 10 December 2024; accepted 25 February 2025.