Good Practices and Common Pitfalls of Machine Learning in Nutrition Research

Daniel Kirk[1*]

[1]Daniel Kirk, King's College London, St Thomas' Hospital Campus, 4th Floor South Wing Block D, Westminster Bridge Rd, London SE1 7EH, United Kingdom, +44 20 7188 5555,

***Corresponding author: Email:** daniel.1.kirk@kcl.ac.uk

Short title: Machine learning: good practices in research

Keywords: Machine learning, data processing, feature selection, reproducible research

DOI 10.1017/S0029665124007638

Proceedings of the Nutrition Society is published by Cambridge University Press on behalf of The Nutrition Society.

Abstract

Machine learning is increasingly being utilized across various domains of nutrition research due to its ability to analyse complex data, especially as large datasets become more readily available. However, at times, this enthusiasm has led to the adoption of machine learning techniques prior to a proper understanding of how they should be applied, leading to non-robust study designs and results of questionable validity. To ensure that research standards do not suffer, key machine learning concepts must be understood by the research community. The aim of this review is to facilitate a better understanding of machine learning in research by outlining good practices and common pitfalls in each of the steps in the machine learning process. Key themes include the importance of generating high-quality data, employing robust validation techniques, quantifying the stability of results, accurately interpreting machine learning outputs, adequately describing methodologies, and ensuring transparency in reporting findings. Achieving this aim will facilitate the implementation of robust machine learning methodologies, which will reduce false findings and make research more reliable, as well as enable researchers to critically evaluate and better interpret the findings of others using machine learning in their work.

Introduction

Machine learning has gained substantial interest in nutritional sciences over the last decade [1]. A PubMed search using the terms “nutrition” and “machine learning” shows the number of articles with title and abstract matches increasing exponentially from 2013 onwards (Figure 1). Examples of applications of machine learning can be seen in various areas of nutrition research, including precision nutrition [2], malnutrition [3], obesity [4], food intake assessment [5], diet recommendation [6] and chatbots for nutritional support [7].

The growing interest in machine learning can be attributed to its appealing properties. Machine learning has the capability to automate tasks that would otherwise be performed manually, thereby freeing up human resources for other activities. Additionally, the different approaches and focuses involved in machine learning compared to traditional statistical methods bring the possibility to analyse data in new ways, which could lead to new scientific discoveries and ultimately improve individual and population health [8,9].

As with any research tool, proper use is essential to ensure the validity of the findings. Unfortunately, the enthusiasm around machine learning has led to adoption preceding a proper understanding of its workings by those applying it [10,11]. This has become apparent in

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

various ways, including the application of machine learning on datasets to which it is not suited, inappropriate methodological choices in data processing steps, non-robust validation schemes, misinterpretation of results, and inadequately described methodology [11]. The consequences of these issues and similar ones include false findings, models that do not generalize to unseen data (i.e., overfitting), and ultimately a reduction in the quality of the literature in the nutrition field.

Claims about the properties of machine learning are used to justify its use, with the considerations behind these claims sometimes neglected. For example, it is often claimed that machine learning techniques are more flexible and make fewer assumptions about the data than traditional statistical methods [12]. However, this does not mean that careful methodological planning and data processing are no longer necessary [9,13]. Even if fewer statistical assumptions are made by some of the algorithms, improper data processing can still lead to suboptimal results.

Machine learning approaches are also praised for their ability to handle high-dimensional datasets [12,14]. For example, ordinary least squares regression cannot be fit when the number of predictors exceeds that of the number of observations because a unique solution to the problem cannot be found (Montgomery et al., 2021). In contrast, machine learning regression algorithms generally allow this without any apparent problem, even though this can lead to overfitting and unstable feature importance estimates, which may go unnoticed unless checked by the analyst [14]. Indeed, the ease with which machine learning experiments can be performed by certain programmes or software libraries and the way in which the outputs are presented can give a false sense of certainty about the results that are generated. Without a better understanding of machine learning and its capabilities and limits, issues will persist in the literature.

To shed light on some of these issues, this review briefly discusses the concept of machine learning before going through steps in the machine learning process and describing good practices and common pitfalls or misconceptions in each. In light of the “new data” theme, there was a focus on concepts of machine learning as it tends to be applied to modern datasets observed in the nutrition sciences, such as large cohorts and omics datasets. The goal was to increase awareness on important details of the machine learning process, promoting robust methodologies in research and enabling a better understanding when interpreting the work of others using machine learning.

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

Machine Learning Overview and Advantages

Machine learning is a subdivision of artificial intelligence (AI) that learns patterns in a dataset to perform a given task without being explicitly programmed to do so [16]. Different types of machine learning exist, within which tasks are performed to achieve an objective.

The two most common types of machine learning in nutrition research are supervised and unsupervised machine learning. In supervised learning, data come with labels and thus the target is known. Either regression or classification are the tasks that are completed to predict the output labels, with algorithms including logistic regression, decision trees, random forest, and support vector machine being used to do this. In unsupervised learning, labels are not available and instead patterns or similarities within the data structure are sought. Tasks include clustering and dimensionality reduction, with example algorithms including k-means clustering and principal component analysis (PCA).

In semi-supervised learning, some of the data (usually a small portion) have labels whereas others (usually a large portion) do not. A combination of both supervised and unsupervised tasks and algorithms may be applied. Reinforcement learning is the final type of learning in which the algorithm updates its behaviour based on feedback from a dynamic environment. Reinforcement learning is currently less often seen in nutrition research but is involved in chatbots and recommendation systems [17–19] and will likely become more common as personalised nutrition grows and chatbots improve. Detailed descriptions of machine learning types and the algorithms used to complete the tasks within them can be seen in Kirk et al.[1].

Machine learning approaches have certain attractive properties which have motivated their inclusion in scientific research methodologies. Being able to learn for themselves how to complete tasks without explicit programming brings the possibility of continuous improvement with the addition of new data [20]. Additionally, the principles of machine learning are not limited to single domains, which means that machine learning can be applied to many different problem areas (provided the data are suitable).

Machine learning can automate the jobs that have historically been undertaken by humans, particularly repetitive ones and those with elements of pattern recognition. One example of this is the research area involved in tracking food intake to improve the accuracy of food intake assessment while simultaneously reducing the burden for those doing so [5]. Many studies in this area make use of machine learning to model unstructured data such as videos of subjects eating [21], pictures of food [22] or audio-based approaches [23]. Machine learning

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

solutions are usually also much faster and can be permanently available, unlike human counterparts that may perform similar duties. For example, ChatGPT not only provided answers to common nutrition questions that scored higher than those from dietitians but could also do so instantly and at any moment of the day [24].

In comparison to traditional descriptive statistical methods, there is usually a focus on prediction on unseen data in machine learning [12]. Hence, in problem areas where prediction is more important than a detailed understanding of the contribution of a set of variables to an outcome, machine learning may be preferable. Unsupervised machine learning can be used for uncovering relationships within complex data structures even in the absence of predefined hypotheses. For example, clustering is often used to group individuals with similar characteristics who might have similar physiological responses to foods or nutritional interventions, such as in metabotyping studies [25]. However, key characteristics of the groups such as which features define them, how many there might be and if they even exist at all is not known beforehand.

Finally, machine learning techniques and traditional statistical methods are sometimes pitted against one another to compare which performs in a certain problem area [26–32]. Whilst such studies might be well-intentioned and simply wish to inform on the effectiveness of a given method, machine learning and traditional statistics should not be thought of rivals competing for the same space. Instead, they should be seen as complementary tools with significant overlap, though also with distinct properties and use cases [12,33,34]. This is perhaps exemplified by techniques which could belong to either category depending on how they are used, such as Lasso [35]. When the goal is inference, where the focus is on drawing conclusions about a population sample and describing underlying relationships within the data, this is a case for traditional statistical methods. When the goal is predictive performance on unseen data, machine learning techniques would be used [12].

Steps in Machine Learning

The machine learning process is composed of a series of steps which start with collecting the data and end with deployment. These steps are described below, although in the context of research, interpreting (**Interpretation**) and describing the methodology and results (**Reporting**) are discussed in place of deployment. At each step, key points, good practices and common misconceptions or pitfalls within them are addressed, as summarised in Table 1.

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

Step 1. Data Collection

Data collection is arguably the most important step in the machine learning process because of its influence in determining the quantity and quality of the data available for modelling [16]. Broadly, data collection can be done either with a research question in mind, for which data are required to investigate, or a dataset in mind, which can be used to investigate a variety of questions.

In the case of the former, it should be ensured that data collected are relevant to the problem and that the collection methods are capable of generating data of a sufficient quality [11]. Whilst machine learning algorithms may be able to learn structures in datasets that are not apparent to humans, they cannot produce meaningful results from poor-quality data (akin to "garbage in, garbage out"). Similarly, machine learning algorithms require an adequate number of instances from which to learn, making sample size an important factor. The number of data points (i.e., individual observations or instances containing unique information) required to achieve adequate performance and reliable results varies depending on data quality, signal to noise ratio, and the machine learning task being performed. In general, however, sample sizes below around 100 are considered small for many supervised and some unsupervised machine learning approaches in nutrition research using biological data and may not provide enough instances from which the algorithms can learn. It is also important that the sample is representative of the population for which the final model is intended. When this is not the case, models may fail to generalise or struggle upon encountering observations with data that were absent in their training (e.g., Naïve Bayes) [36].

Alternatively, data may also be collected without a specific research question in mind and where the goal is to create a dataset of sufficient size and depth to answer a broad array of questions and remain relevant over a long period [37]. It remains important that the sample is representative of the population of interest, meaning inclusion criteria must be carefully thought out [38]. For example, the inclusion and exclusion criteria for a large cohort study must ensure that participants are representative of the target population and recruitment techniques must be selected in a way that minimises biases [39]. Various types of data should be collected to permit investigating questions on a broad range of topics and the methods used in their collection should be documented in detail [40]. Important considerations include the longevity of the data collected, questionnaire wording and response options, which data (variables) will

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

be collected, data storage (both physically and digitally), privacy and ethical considerations, and documentation and metadata, amongst others.

There has been much focus in recent decades on improving machine learning algorithms or developing new ones [40]. Great strides have been made in this regard, and there are now many algorithms available for various machine learning tasks and problem areas [41]. Despite this, data quality remains the most important limiting factor on the performance of most machine learning applications. It is unlikely that future breakthroughs will occur solely through the development of new and improved algorithms; rather, there must be a focus on improving data quality through rigorous methods of data collection and processing (discussed below) [40].

Step 2. Data Processing

Once collected, data usually require processing. The methodological decisions made during data processing influence the data that is eventually used for modelling. This section describes common data processing steps that should be considered in a machine learning experiment. Importantly, some data processing steps should be performed within validation steps and not applied to the whole dataset in order to avoid information leakage (see below: ***Internal validation schemes*** in **Step 3. Modelling**). Attention is brought to these situations below.

Selecting observations

It is possible that not all of the observations in the dataset are suitable to be included in the analysis. Reasons for this could include outliers, repeated measures (from which only one is required) or subgroups with few observations, such as males in a predominantly female sample. Decisions regarding which observations should be included and excluded should justified (e.g., based on good statistical grounds or findings from previous studies) and well documented when the methodology is described.

Pre-selecting features & dimensionality reduction

Whilst classic feature selection makes use of the outcome and uses statistical techniques to determine which features to include (see below: **Step 3. Modelling**), the analyst may also have to decide which features (or groups of features) are potentially relevant to the problem and therefore should be included for data processing. Where possible, domain knowledge

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

should be used to justify the elimination of variables that are not expected to be relevant to the problem [9,42]. For example, when investigating cardiovascular disease risk using a large cohort, specific biochemical measures may be included, whereas others deemed insufficiently relevant are excluded. Importantly, any feature selection steps that make use of the outcome variable must be performed within validation loops (see below: ***Internal validation schemes*** in **Step 3. Modelling**).

Dimensionality reduction refers to reducing the number of features irrespective of the outcome variable (i.e., unsupervised approaches), such as through identifying redundancies in the data [43,44]. Reducing the dimensionality of a dataset can be desirable by reducing model complexity, computation time, problems related to collinearity, and overfitting [42]. Low-variance features may not contain enough information to be useful to the problem and are sometimes removed or combined with other features, if appropriate [16]. However, this should be done with caution since loss of information is possible. For example, in microbiome studies it is common to see bacteria present in fewer than a given proportion of the sample removed (e.g., <5%). However, in some cases the presence of a given bacterium in a small group of individuals may be informative for healthy outcomes or the problem being investigated. Such findings may be lost due to low-variance filtering.

Techniques for identifying similarities or redundancies within the data can also be employed, such as correlation-based approaches, PCA and variable clustering [45,46]. It is commonly believed that it is not necessary to perform such unsupervised data reduction techniques within validation steps and, instead, that they can be safely applied to the entire dataset. This belief is based on the understanding that these techniques do not make use of the outcome variable, thus minimizing the risk of information leakage [14,16]. However, recent findings suggest that unsupervised dimensionality reduction steps on the whole dataset can still introduce bias [47]. Whilst more work is needed on the topic, analysts may consider repeating analyses where dimensionality reduction is confined within training splits of validation steps to assess the sensitivity of results to the timing of these steps.

Processing missing data

Missing data is common in many datasets and comes in different forms, including missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) [48]. Each reflects different underlying mechanisms: MCAR implies that the missingness is unrelated to any data, MAR suggests that the missingness is related to

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

observed data but not the missing data themselves, and NMAR indicates that the missingness is related to the unobserved data [48]. These distinctions are crucial because they influence the methods used to handle missing data and the conclusions derived from the results.

One approach to missing data is to simply restrict the analyses to complete cases. This may be justified in certain circumstances, such as when data are MCAR and the number of observations with missing data is relatively small, however, when data are NMAR, removing observations with missing data can bias the results, and verifying the type of missing data in question is not always possible [49].

An alternative is to impute the missing data in various ways. Single imputation approaches with the mean, mode or other summary statistics are simple to implement, although they are almost never optimal and can distort the distribution of the features of the dataset and the relationships between them [50,51]. In contrast, multiple imputation uses distributions of the observed data to provide multiple plausible complete datasets, thus accounting for uncertainty in the missing values [49]. Analyses are repeated on each dataset and the results are pooled to account for the variability due to imputation.

Finally, model-based approaches, such as regression and machine learning techniques (e.g., neural networks, k-nearest neighbours and random forest) use the observed data to predict missing values [52,53]. Various packages exist which facilitate the implementation of imputation techniques [54]. Imputation of missing data should be done within validation steps.

Processing outliers

Outliers can arise due to a variety of reasons, including equipment malfunction, human error during data entry, or extreme (but valid) data measurements [55]. Regardless of their origin, the presence of outliers should be investigated during exploratory data analysis and, if necessary, appropriate action should be taken to account for them [9]. Outliers can sometimes be detected through manual observation of features through descriptive characteristics or plots [56]. For example, if the median LDL cholesterol levels in a sample were 2.5mmol/L, with an interquartile range of 1mmol/L but a maximum value of 25mmol/L, it would be likely that one or some values were unrealistically high, possibly reflecting a mistake during manual entry or measuring equipment malfunction. Other times, however, what defines an outlier is less clear, and even in cases where extreme values are found, it is not always the case that some type of treatment is required.

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

Whilst general approaches based on statistical properties can simple and easy to implement (e.g., values greater than the third quartile plus 1.5 *interquartile range) [55], they often do not consider the characteristics of the data or the problem at hand and are unjustified in many situations. For example, a small number of people may report a much higher income than the rest in a sample; however, these may be perfectly valid observations that do not require corrective action, despite their appearance in descriptive statistics or plots. Whenever possible, domain knowledge and additional related information should be used to guide decision-making for identifying and processing outliers, and this should be documented in the methodology.

If observations are determined to be outliers, if and how they should be dealt with should be based on logical grounds and on a case-by-case basis, whenever possible. When this is not possible, such as when features cannot be easily interpreted or in high-dimensional settings, identifying and dealing with outliers should be done carefully to avoid loss of information or risk of introducing bias.

Finally, the processing of outliers may be done on either the whole dataset or within validation steps, depending on the purpose. For instance, approaches to correcting genuine errors in the dataset, such as mistaken data entries, can be performed on the whole dataset as there is little risk of introducing bias. However, if characteristics of the data are used to identify and correct potential outliers, this should be done within validation steps.

Transforming features

Feature transformation, such as normalisation or standardisation, is another common data processing step and may be required by some algorithms when the features are on different scales. Examples include the regularized regression techniques Lasso and Ridge regression, where the scale of the features is relevant to the penalty term [35], and k-nearest neighbours, where the scale of the features is relevant is relevant to distances calculations [57]. Whether transformation is required for the algorithms used for data analysis should be known by the analyst beforehand. However, it should be noted that the choice of transformation technique used can have a significant impact on the results and should not be an arbitrary decision[58–60]. Hence, if transformation techniques are required, it may be warranted to investigate how the results change with different transformation techniques in order to assess the sensitivity of the findings. Transformations such as normalisation should be performed within validation steps to prevent data leakage.

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

Discretisation of continuous features

This refers to converting features on a continuous scale to categorical ones. A common example in nutrition is the conversion of BMI in kg/m^2 to underweight, healthy weight, overweight and obese. Despite the dangers of this practice being well-documented and long-known [43,61–65], it remains widespread in the literature.

Discretisation almost inevitably leads to a loss of information, which can hinder the predictive capacity of the model [61,64]. The fewer the number of levels (or bins) in the newly formed category, the greater the loss of information. Another consequence of discretisation is the introduction of step functions where the response "jumps" when moving from one level to the next within a category [43]. Unless this is justified (e.g., the pH at which a biological reaction occurs), such situations are usually undesirable and, in biological settings, may be unrealistic [65].

The decision regarding the number of bins to use ("binning") and the numerical limits which define them is also problematic [64]. This is sometimes done in terms of quantiles (e.g., quartiles) or groups that make sense to humans (e.g., age groups of 40-49 years, 50-59 years, etc.), despite that such groups may not make sense to the problem at hand [62,63]. An additional problem related to binning concerns the sensitivity of the results to the limits defining the bins. In the absence of well-defined cut-points based on prior work [61], this decision is left in the hands of the analyst. Unintentionally or otherwise, this can lead to cut-points being selected that support hypotheses or iteratively trying enough combinations of bins and cut-points until favourable results are found. The consequences of such practices include an increased chance of spurious findings, false positives and biased results [61,65,66].

Finally, discretisation of the outcome variable can influence the information that can be obtained from the results. Vastly different observations may be grouped together, whereas observations with only minimal differences may end up in different levels [61,63]. For example, discretising blood pressure measurements into levels of hypertension (e.g., normotensive, pre-hypertension and hypertension) could lead to individuals with dangerously high blood pressure and those who are barely hypertensive being classified into the same group, whereas those only 1 mmHg apart could be assumed to have different risks [61]. This can have important consequences for the conclusions derived from the results, resource allocation and treatment or intervention options.

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

Step 3. Modelling

The algorithms suitable for a specific problem depend on the machine learning task it involves. Various options exist for regression, classification, clustering, dimensionality reduction, and reinforcement learning, and some algorithms can be used for multiple tasks [41].

The myriad options available can be overwhelming, and it is usually difficult to deduce on purely theoretical grounds which methods will perform best on a given problem beforehand [16]. For this reason, it is common to implement different ones and compare their performances [67]. One important consideration in doing this is making sure that hyperparameter tuning (discussed below) takes place on a level playing field. Failure to ensure this risks providing an unfair representation of the results since some algorithms may perform comparatively well with little or no tuning (e.g., random forest [68]), whereas others can be more sensitive to this. Similarly, the data used for training and testing should be the same for each candidate algorithm.

Sometimes the performance between different models may differ only slightly and be of little practical relevance. Whilst a given model may achieve the best performance in a given experiment, a certain amount of variability should always be expected and the possibility that the performance of the other models used for comparison could have been different had a different data sample been used cannot be ruled out. Hence, analysts should avoid overemphasising small performance differences and instead set an acceptable limit of tolerance (determined in the context of the problem and possibly set in advance and described in the methodology) within which various models could be considered. If formal testing is preferred, statistical tests can also be used to compare the performance of different machine learning models [69].

Prediction quality on scoring metrics, such as accuracy or mean squared error (see below: **Step 4. Evaluation**), is an important factor in selecting a machine learning model, but it is not the only one. Interpretability and calculation speed can also motivate the selection of machine learning algorithms. For example, ensemble methods are a group of algorithms that combine the results of many individual learners as their final output [70], with notable examples including random forest[71] and XGBoost[72]. By aggregating multiple individual learners, ensemble methods are less sensitive to the errors that individual learners make and thus tend to have lower variance and, in general, perform well on unseen data compared to methods which only rely on single learners [73]. However, this comes at the cost of longer fitting times

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

and lower interpretability, which may motivate the selection of simpler methods (even if predictive accuracy is lower).

Hyperparameter tuning

Hyperparameters are modifiable parameters that affect the learning process of an algorithm. Examples include maximum depth on decision trees and the learning rate in neural networks. How these hyperparameters are tuned can have a significant impact on model performance. The tuning process involves fitting many different models, each with different hyperparameter configurations, to see which set of configurations leads to the best performance. Grid search and random search [74] are relatively basic optimization techniques and are commonly implemented in nutrition research literature, though more advanced techniques also include Bayesian Optimization, Hyperband, and evolution strategies [75].

Feature selection

Feature selection aims to reduce the initial feature space by eliminating features which less important to the model output, ideally whilst preserving predictive performance [76]. This is increasingly sought-after as data in the modern world are being collected on a wide range of variables, sometimes from only a small number of samples. For example, thousands of microbes can be collected from each individual in gut microbiome studies, yet such studies rarely have so many participants.

Many feature selection approaches exist [77,78], although the difficulty of the task of feature selection in the high-dimensional setting is often underappreciated [79]. Especially when the number of features is much higher than the number of observations, identifying features which generalise to unseen data and distinguishing those that are relevant to the problem from those that are not can be challenging [79–81].

An additional challenge is feature selection stability, which refers to how sensitive the selected features are to perturbations of the data [82]. Ideally, the feature subset would contain features that are selected across multiple repeats of feature selection, each using different splits of the data for training and testing. Feature selection stability can be estimated by repeating feature selection across multiple different subsamples of the data, as part of robust internal validation schemes [81,82] (as described below).

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

Internal validation schemes

Many techniques for hyperparameter tuning, feature selection and data processing steps make use of the outcome variable, meaning there is risk of information leakage [83,84]. Information leakage describes the situation in which the same observations in a dataset are used to both construct and evaluate the model, resulting in an optimistic evaluation of model performance on unseen data [83,84].

To minimise information leakage, robust internal validation schemes preserve an entirely unseen portion of the data that was not involved in data processing or model-building for validation. Various approaches exist for this and are discussed below (see below: **Step 4. Evaluation**), though in general they involve splitting the data for training, where data processing, hyperparameter optimisation and feature selection are performed, and testing, where evaluation is performed. This procedure is then repeated multiple times to estimate uncertainty and stability in the results [85]. This can protect against spurious findings and make the results more robust. A visual representation of how a robust internal validation scheme might look can be seen in Figure 2. Examples of machine learning experiments with good internal validation schemes can be seen in [86,87].

Step 4. Evaluation

The purpose of model evaluation is to assess how well the model has performed its task, usually with a focus on performance on unseen data. Internal validation approaches divide the dataset into different portions for model-building and evaluation and would, ideally, be followed by external validation, where the models developed on the original dataset are validated on an unrelated dataset [1]. Below, some common validation techniques are introduced, along with metrics by which optimisation occurs and model performance is evaluated.

Metrics

The metrics by which machine learning modes are evaluated are important since they not only reveal how a model performs from different perspectives but also determine how they are optimised [88]. Hence, it is important that metrics are chosen that reflect the desired performance of the models with respect to the problem [11]. In some cases, performance with respect to some metrics may be more important than others. For example, in a classification problem, it may be more important that all true cases are correctly identified even at the

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

expense of incorrectly labelling true negatives as positives (i.e., high sensitivity, low specificity), such as in identifying children at risk of obesity for targeted advertisement for events at a local sports centre. In other circumstances, false positives can be costly, such as when cases predicted as positive are selected to undergo surgery. In this case, a low specificity could lead to unnecessary surgery and related complications. Alternatively, if there is no particular preference for one metric over others, composite metrics which consider multiple aspects of performance (e.g., F1 score [88]) may be preferred.

Similar differences exist for metrics in regression. For example, mean-squared error (MSE) and root MSE (RMSE) punish larger errors more than mean absolute error (MAE), which can be desirable in some cases [89]. The coefficient of determination, R-squared, on the other hand, measures the proportion of the variation in the outcome that is explained by the features [90]. Because R-squared is limited between 0 and 1, it is easier to compare performance between different datasets that may have different variables on different scales [90]. However, it should be noted that it is not always the case that optimal performance with respect to one single metric is desired; instead, models may be evaluated across different metrics with each evaluating different aspects of performance [90].

Whereas in supervised machine learning labels are available for the data, providing a ground truth on which to score predictions, this is not the case with unsupervised machine learning, which uses different metrics for assessing model performance. The exact metrics depend on the specific unsupervised task and often involve measures of homogeneity or dissimilarity. For example, for clustering, one of the most common unsupervised machine learning tasks, metrics include silhouette score, Calinski-Harabasz coefficient, and Dunn index [91]. The different ways in which these metrics determine the quality of clustering can lead them to arrive at different optimal solutions. Hence, unless there is rationale to prefer one over another, it may be interesting in unsupervised approaches to explore how results differ across different metrics and how this influences the eventual conclusions drawn from the machine learning experiment.

Validation procedures

Train-test split

Splitting the dataset into a portion for training (model-building) and a portion for testing (evaluating model performance) is the most basic way in which a model can be evaluated [1].

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

The data may be split entirely randomly or in a way that maintains certain aspects of the data characteristics in each portion, such as ensuring the same proportion of cases and controls in the training and test data. The cost of this simplicity is that the results obtained from a train-test split can be highly dependent on how the dataset was split, which can lead to significant variability in performance metrics, especially with smaller datasets. While train-test splits are more reliable on large datasets or those with good signal-to-noise ratios, alternative methods such as cross-validation (discussed below) offer a more robust evaluation of model performance. If train-test splits are to be used, they should be repeated in order to increase the stability of the results [92].

The importance of repeating validation techniques with different subsamples of the data is shown in Figure 3, which uses a decision tree classifier with the Pima Indians Diabetes dataset [93] to demonstrate the effect of validation technique, the number of times it is repeated, and sample size on the stability and uncertainty of the results. The code for the analysis and location of the dataset can be seen in the supplementary code. Figure 3 makes clear the importance of repeating validation techniques on smaller datasets in order to increase the certainty of the results. However, this is not always observed in the literature, despite the fact that relatively small samples are often used in machine learning experiments.

Cross-validation

Cross-validation approaches are a group of resampling procedures that can be used for model selection and to estimate the generalisability of the model [14]. In *k*-fold cross-validation, the dataset is split into *k* folds so that each observation is used once for testing and *k*-1 times for training [14]. The number of folds *k* is often chosen based on computational efficiency and a suitable bias-variance trade-off, with values generally between 5 and 10 being used in practice [92]. Different variations of cross-validation exist, such as leave-one-out cross-validation and stratified and grouped cross-validation (see Kirk et al.[1]). Unlike train-test split, there are *k* number of test score results, which may be presented individually (as in Figure 3) or aggregated into summary statistics (e.g., mean across all test folds). Different varieties of cross-validation exist to allow stratified cross-validation or account for dependent observations.

Cross-validation is an improvement over train-test split because all of the data are used for training and testing, making it less sensitive than a single split in the data. Even so, cross-validation can still be sensitive to how the data is split within each fold and test scores

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

between each fold can differ greatly, especially with smaller datasets. To account for this, cross-validation can be repeated multiple times over, ensuring that on each repeat different splits of the data are used within each fold (Figure 2) [85,92].

Nested cross-validation

One concern with cross-validation is that data processing steps, hyperparameter tuning and feature selection are performed on the same data used to evaluate the model performance. Nested cross-validation deals with this by adding another cross-validation loop (known as the inner loop) within the training data of each fold of the outer loop (see Figure 2). The inner loop is used for tuning hyperparameters and feature selection, and this is then validated on the test data of the outer loop (for details see [94]). This ensures a portion of the data that was not involved in any part of the model-building process is available to estimate performance on unseen data.

Nested-cross validation reduces the chance of information leakage and allows for an unbiased estimation of model performance [95]. However, in doing so it greatly increases computational time, and whether this justifies the reduction in bias has been called into question [94]. Analysts may prefer to first perform traditional cross-validation and then, if the results appear promising, validate that the findings are not the result of optimism due to information leakage by using nested cross-validation. This can circumvent wasting time and computer resource on machine learning experiments that would not be fruitful anyway.

Calibration

An important yet often overlooked concept for supervised classifiers is calibration, which refers to the alignment between the predicted probabilities and the observed outcomes [96]. For example, a classifier that is trained to predict diabetes would be well-calibrated if the observed occurrence of diabetes was close to 10% for those whose predicted probabilities were close to 10%. Model miscalibration may not always be apparent in internal validation, but low calibration can mislead expectations and be problematic when looking at high or low-risk groups (where it may be needed most) or external datasets [97]. This can have important implications for the action taken in response to the predictions made by the model.

In contrast to predicting class labels directly, there are advantages to working with predicted probabilities [97]. Firstly, it is often more relevant to know the chance of an event occurring rather than a class label without further context. For example, whilst two individuals may

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

both be predicted to belong to the same class, the probability for one being 55% and the other being 95% shows a clear difference in the confidence of their predicted membership. This can have important implications, such as how resources are allocated in response to model predictions (e.g., children with a higher predicted probability of malnutrition are prioritised for corrective nutritional intervention). Using probabilities also means that custom thresholds can be more easily set and adjusted, which can be desirable when the cost or benefit of correct classification is not the same as that for misclassification. Finally, probabilities are also inherently more interpretable than class labels, which is a desired property of machine learning procedures [98].

Miscalibration may occur due to the data or the model used to fit the data (i.e., overfitting) [97]. It is most often diagnosed by plotting the predicted probabilities against the observed frequencies (known as calibration curves or reliability plots), with a straight-line $y=x$ representing perfect calibration [99]. Due to differences in how they operate, some models drive probabilities away from 0 and 1 (e.g., support vector machines), whereas others more readily predict probabilities at the extremes (e.g., Naïve Bayes) and others still are naturally well-calibrated (e.g., logistic regression) [100,101]. Following the identification of miscalibration, calibration correction techniques can be used. Two well-known approaches are Platt scaling [102], which is used for those with S-shaped calibration curves, such as those seen for support vector machines, and isotonic regression, which is capable of modelling more complex shapes but with an increased risk of overfitting [101].

External validation

A primary goal of machine learning applications is to make predictions using new data that were not available during training. While good internal validation schemes provide an estimate of how well prediction models do this, they can still be optimistically biased [11,103]. External validation involves validating the generalisability of the model on a dataset that reflects the target population but comes from another source [11,96,103–107].

External datasets may differ in key characteristics such as the location, time and methods of data collection, as well as the individuals responsible for collecting the data. However, it is crucial that key features remain constant. For instance, when externally validating a model predicting disease outcomes, it is imperative that the disease is defined in the same way in both datasets to prevent differences in model performance reflecting disease definitions rather than generalisability. To further enhance robustness and reduce the chance of optimistic bias,

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

external validation may be performed by an independent research group that was not involved in model development or data collection for the initial dataset [11].

It is worth noting that some studies sometimes erroneously claim that external validation was used, when in fact their "external validation" set is simply a test set or an extension of the original dataset (e.g.,[108]). This is sometimes a semantic issue rather than intentional misrepresentation [107], but care should be taken when interpreting such results because external validation is a stronger sign of generalisability than internal validation with a larger dataset. Even still, it should be kept in mind that good external validation performance does not prove generalisability [105,109,110]. External validation performance is still dependent on the external dataset used and conclusions made must bear the characteristics of this dataset (e.g., location, time, sample characteristics, etc.) in mind. As external validation is repeated on a greater number of external datasets that are more different from the dataset used to develop the model, confidence in the generalisability of the model increases [109,110].

Step 5. Interpretation

Metrics by which machine learning models are optimised were introduced in **Step 4. Evaluation**, though it is also important that they are correctly interpreted and presented. For example, the simplest and most common metric for evaluating classifier performance is accuracy; however, accuracy can be deceptive. This is particularly evident in imbalanced datasets (i.e., where the proportion of one class is much higher than the other), where a classifier which always predicts the majority class (irrespective of the data it receives) can score highly on accuracy, despite having no predictive capacity [88]. Similar considerations exist for other metrics, such as MSE for regression, which is dependent on the scale of the outcome variable, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), for which scores of around 0.65 are sometimes viewed positively, despite 0.5 being what could be expected with random guessing.

The interpretation of machine learning results can be made complicated when using complex validation schemes. For example, in comparison to a single train-test split, in which metrics from one portion of test data can be easily understood and reported, machine learning experiments which involve multiple repeats of train-test split or cross-validation can have scores from many test sets which may need to be summarised concisely to be understood. Summary statistics on the results can be useful, such as reporting the mean, median, range and interquartile range, provided there are enough test scores for such statistics to be

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

meaningful. Plots of results, such as those seen in Figure 3, can also be useful to present many results at once without loss of information.

Feature importance

It is often desirable to know which features were important for machine learning models in generating their output, despite that there is usually no ground truth for feature importance and the concept itself is poorly defined [111]. Different approaches estimate feature importance in different ways, often arriving at different conclusions [112], though popular approaches with good statistical properties include SHAP [113], Lime [114], and permutation-based feature importance [115], amongst others [111].

Some algorithms can provide feature importance estimates as part of their architecture (so-called "built-in" feature importance). Such built-in feature importance estimates, whilst convenient, can come with significant drawbacks. This is particularly well-described for random forest [71], which can provide biased results based on the scale of continuous features and number of categories in categorical ones, as well as when multicollinearity is present [116–118]. While no feature importance method is perfect [111], it is concerning how often random forest-derived feature importances are reported in scientific literature without consideration for their potential limitations or how the results may be different had other feature importance techniques been used [2], creating a false sense of certainty for features importance estimates calculated with this approach.

Inappropriately calculated feature importance estimates can lead to both false positives (i.e., unimportant features falsely identified as important) and false negatives (i.e., important features falsely identified as unimportant), both of which reduce the quality of the literature. In response to this, analysts should first think carefully about how their choice of feature importance technique relates to their data [112]. It should be known if there are feature interactions or collinearity present, and how the chosen feature importance techniques may be affected by this [111]. Analysts should also be open to comparing results across various suitable techniques and place their findings in the context of the model and explainable AI technique used [119], rather than making conclusions about feature importance in a general sense.

Additionally, an important but sometimes underappreciated fact is that feature importance estimates on models with poor predictive performance cannot be trusted, and analysts should resist reporting feature importance estimates in such cases [112,120]. Moreover, similarly to the

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

results from predictions generated by machine learning models, feature importance estimates may also be dependent on how the dataset was split for training and testing [121,122]. It can be informative to estimate the stability of feature importances by seeing how they change across many repeats of model fitting on different splits of the data [111,121,122]. Finally, the subfield of explainable AI is not without issues and has been criticized for the consequences of unfaithful explanations, complex explanations, failing to consider that important features may change over time, and ambiguity regarding terminology [123–125]. Hence, researchers should be aware of the shortcomings of explainable AI techniques they decide to use and the potential costs of these in the context of the problem.

Step 6. Reporting

Reporting refers to describing both the methodology of the machine learning experiment and reporting the results obtained. There have been numerous concerns raised about the reproducibility of published work in the health field [126–129] and, unfortunately, the growing use of machine learning in research may further exacerbate this issue [130]. This is owed in part to difficulties involved in the machine learning process, such as minor details in data processing, modelling, and evaluation that may go undocumented, along with other factors such as randomness, software versions, data availability, biased methodologies, and selective reporting of optimistic results [130,131]. Because of this, it is crucial that studies using machine learning in their research describe their methodology and report their results in appropriate detail.

The methodology in studies using machine learning should be described as thoroughly as possible, in a way that another analyst would be able to obtain the same results if they had access to the same dataset [131,132]. In this regard, publishing code can be helpful because most, if not all, steps of the analysis can be automatically documented within this. Another advantage of publishing code is that it may still be interpreted and understood even without access to the data, and if the data is made available at some point in the future, investigating reproducibility is made much easier. When code is not available it becomes more important that each step, in their correct order, is described in adequate detail.

It is not uncommon to see data processing steps described in insufficient detail [132]. For example, if there were any missing data or outliers present in the initial dataset, their processing should be described in sufficient detail to allow an external party to perform exactly the same steps on the relevant data points. If there were no missing data outliers, then

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

this should be mentioned, otherwise, it can be inferred that some level of data processing occurred that was not documented, which brings into question the validity of the whole experiment.

The same applies to many other steps in the machine learning process, such as optimising hyperparameters, model validation schemes and feature selection. To ensure reproducible work, these and other steps should be described in adequate detail with respect to the complexity of the procedure, with accompanying random states or random seeds provided [131]. Statements such as “hyperparameters were optimised” or “cross-validation was used” should not be acceptable without further specification of how this was done.

When reporting the results of machine learning experiments, it is important to be transparent about the findings [10,85]. It can be tempting to report only those results that make the experiment seem successful, such as reporting only favourable results and ignoring those which make the findings seem less convincing [85]. However, all findings can be interesting and may provide different information, which can be useful for informing future work based on the results. For example, the consequence of omitting findings which expose instability in the results could be that money and time are wasted on validation studies. It is particularly important that any changes to the data processing or modelling steps that were informed by the outcome variable are documented, otherwise, results can be become biased, similar to p-hacking with traditional statistical methods [133].

Conclusion

Machine learning has the potential to be a valuable tool in nutrition research. For this potential to be realised, it is imperative that researchers have sufficient understanding of machine learning concepts to be able to interpret the results of others and apply well-designed machine learning methodologies themselves. Failure to achieve this will lead to a reduction in the quality of the literature, missed opportunities, and wasted resources in unproductive efforts to validate or extend upon existing work.

By going through each of the key steps in the machine learning process, this review and the conference proceedings which it documents aimed to provide an overview on good practices and highlight common misconceptions and pitfalls of using machine learning in nutrition research. Nutrition researchers using machine learning in the coming years should focus on the generation of high-quality data, robust validation techniques, quantifying the stability or

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

uncertainty of results, proper interpretation of machine learning outputs, adequately described methodologies, and transparency when reporting results.

Acknowledgements

I would like to express my gratitude to the Nutrition Society, particularly Dr Anne Nugent and Professor Jayne Woodside, for inviting me to speak at the inaugural Nutrition Society Congress, which was an excellent event. I would like to thank also Andrea Onipede for her organisational support before and during the congress.

Financial Support

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Declaration of Interests

Declaration of interests: The author declares none.

References

1. Kirk, D., Kok, E., Tufano, M., Tekinerdogan, B., Feskens, E.J.M., and Camps, G. (2022). Machine Learning in Nutrition Research. Advances in Nutrition *13*, 2573. https://doi.org/10.1093/ADVANCES/NMAC103.

2. Kirk, D., Catal, C., and Tekinerdogan, B. (2021). Precision nutrition: A systematic literature review. Comput Biol Med *133*. https://doi.org/10.1016/J.COMPBIOMED.2021.104365.

3. Janssen, S.M., Bouzembrak, Y., and Tekinerdogan, B. (2024). Artificial Intelligence in Malnutrition: A Systematic Literature Review. Adv Nutr. https://doi.org/10.1016/J.ADVNUT.2024.100264.

4. DeGregory, K.W., Kuiper, P., DeSilvio, T., Pleuss, J.D., Miller, R., Roginski, J.W., Fisher, C.B., Harness, D., Viswanath, S., Heymsfield, S.B., et al. (2018). A review of machine learning in obesity. Obes Rev *19*, 668–685. https://doi.org/10.1111/OBR.12667.

5. Oliveira Chaves, L., Gomes Domingos, A.L., Louzada Fernandes, D., Ribeiro Cerqueira, F., Siqueira-Batista, R., and Bressan, J. (2023). Applicability of machine learning techniques in food intake assessment: A systematic review. Crit Rev Food Sci Nutr *63*, 902–919. https://doi.org/10.1080/10408398.2021.1956425.

6. Shah, M., Degadwala, S., and Vyas, D. (2022). Diet Recommendation System based on Different Machine Learners: A Review. Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy, ICAIS 2022, 290–295. https://doi.org/10.1109/ICAIS53314.2022.9742919.

7. Oh, Y.J., Zhang, J., Fang, M.L., and Fukuoka, Y. (2021). A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. International Journal of Behavioral Nutrition and Physical Activity *18*, 1–25. https://doi.org/10.1186/S12966-021-01224-6/TABLES/4.

8. Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. (2022). AI in health and medicine. Nature Medicine 2022 28:1 *28*, 31–38. https://doi.org/10.1038/s41591-021-01614-0.

9. Badillo, S., Banfai, B., Birzele, F., Davydov, I.I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., and Zhang, J.D. (2020). An Introduction to Machine Learning. Clin Pharmacol Ther *107*, 871. https://doi.org/10.1002/CPT.1796.

10. Lipton, Z.C., and Steinhardt, J. (2019). Troubling Trends in Machine Learning Scholarship. Queue *17*. https://doi.org/10.1145/3317287.3328534.

11. Vollmer, S., Mateen, B.A., Bohner, G., Király, F.J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K.S.L., Myles, P., et al. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ *368*. https://doi.org/10.1136/BMJ.L6927.

12. Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. Nat Methods *15*, 233. https://doi.org/10.1038/NMETH.4642.

13. Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics 2015 16:6 *16*, 321–332. https://doi.org/10.1038/nrg3920.

14. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning 2nd ed. (Springer New York) https://doi.org/10.1007/978-0-387-84858-7.

15. Montgomery C, D., Peck A, E., and Vining, G.G. (2021). Introducing To Linear Regression Analysis 6th ed. (Wiley).

16. Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci *2*, 1–21. https://doi.org/10.1007/S42979-021-00592-X/FIGURES/11.

17. Theodore Armand, T.P., Nfor, K.A., Kim, J.I., and Kim, H.C. (2024). Applications of Artificial Intelligence, Machine Learning, and Deep Learning in Nutrition: A Systematic Review. Nutrients *16*. https://doi.org/10.3390/NU16071073.

18. Liu, L., Guan, Y., Wang, Z., Shen, R., Zheng, G., Fu, X., Yu, X., and Jiang, J. (2024). An interactive food recommendation system using reinforcement learning. Expert Syst Appl *254*, 124313. https://doi.org/10.1016/J.ESWA.2024.124313.

19. Yau, K.L.A., Chong, Y.W., Fan, X., Wu, C., Saleem, Y., and Lim, P.C. (2023). Reinforcement Learning Models and Algorithms for Diabetes Management. IEEE Access *11*, 28391–28415. https://doi.org/10.1109/ACCESS.2023.3259425.

20. Jordan, M.I., and Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. Science *349*, 255–260. https://doi.org/10.1126/SCIENCE.AAA8415.

21. Tufano, M., Lasschuijt, M., Chauhan, A., Feskens, E.J.M., and Camps, G. (2022). Capturing Eating Behavior from Video Analysis: A Systematic Review. Nutrients 2022, Vol. 14, Page 4847 *14*, 4847. https://doi.org/10.3390/NU14224847.

22. Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., and Ma, Y. (2016). DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) *9677*, 37–48. https://doi.org/10.1007/978-3-319-39601-9_4.

23. Kalantarian, H., and Sarrafzadeh, M. (2015). Audio-based detection and evaluation of eating behavior using the smartwatch platform. Comput Biol Med *65*, 1–9. https://doi.org/10.1016/J.COMPBIOMED.2015.07.013.

24. Kirk, D., Van Eijnatten, E., and Camps, G. (2023). Comparison of Answers between ChatGPT and Human Dieticians to Common Nutrition Questions. J Nutr Metab. https://doi.org/10.1155/2023/5548684.

25. Palmnäs, M., Brunius, C., Shi, L., Rostgaard-Hansen, A., Torres, N.E., González-Domínguez, R., Zamora-Ros, R., Ye, Y.L., Halkjær, J., Tjønneland, A., et al. (2020). Perspective: Metabotyping—A Potential Personalized Nutrition Strategy for Precision Prevention of Cardiometabolic Disease. Advances in Nutrition *11*, 524–532. https://doi.org/10.1093/ADVANCES/NMZ121.

26. Chowdhury, M.Z.I., Leung, A.A., Walker, R.L., Sikdar, K.C., O'Beirne, M., Quan, H., and Turin, T.C. (2023). A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a Canadian population. Sci Rep *13*. https://doi.org/10.1038/S41598-022-27264-X.

27. Shin, S., Austin, P.C., Ross, H.J., Abdel-Qadir, H., Freitas, C., Tomlinson, G., Chicco, D., Mahendiran, M., Lawler, P.R., Billia, F., et al. (2021). Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. ESC Heart Fail *8*, 106–115. https://doi.org/10.1002/EHF2.13073.

28. Premsagar, P., Aldous, C., Esterhuizen, T.M., Gomes, B.J., Gaskell, J.W., and Tabb, D.L. (2022). Comparing conventional statistical models and machine learning in a small cohort of South African cardiac patients. Inform Med Unlocked *34*, 101103. https://doi.org/10.1016/J.IMU.2022.101103.

29. Panaretos, D., Koloverou, E., Dimopoulos, A.C., Kouli, G.M., Vamvakari, M., Tzavelas, G., Pitsavos, C., and Panagiotakos, D.B. (2018). A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): the ATTICA study. Br J Nutr *120*, 326–334. https://doi.org/10.1017/S0007114518001150.

30. Choi, S.G., Oh, M., Park, D. –H, Lee, B., Lee, Y. ho, Jee, S.H., and Jeon, J.Y. (2023). Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. Sci Rep *13*. https://doi.org/10.1038/S41598-023-40170-0.

31. Belsti, Y., Moran, L., Du, L., Mousa, A., De Silva, K., Enticott, J., and Teede, H. (2023). Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. Int J Med Inform *179*. https://doi.org/10.1016/J.IJMEDINF.2023.105228.

32. Suzuki, S., Yamashita, T., Sakama, T., Arita, T., Yagi, N., Otsuka, T., Semba, H., Kano, H., Matsuno, S., Kato, Y., et al. (2019). Comparison of risk models for mortality and cardiovascular events between machine learning and conventional logistic regression analysis. PLoS One *14*. https://doi.org/10.1371/JOURNAL.PONE.0221911.

33. Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. Front Neurosci *11*, 273651. https://doi.org/10.3389/FNINS.2017.00543/BIBTEX.

34. Breiman, L. (2001). Statistical Modeling: The Two Cultures. Statistical Science *16*, 199–231. https://doi.org/10.1214/SS/1009213726.

35. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) *58*, 267–288. https://doi.org/10.1111/J.2517-6161.1996.TB02080.X.

36. Wickramasinghe, I., and Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft comput *25*, 2277–2293. https://doi.org/10.1007/S00500-020-05297-6/FIGURES/2.

37. Setia, M.S. (2016). Methodology Series Module 1: Cohort Studies. Indian J Dermatol *61*, 21. https://doi.org/10.4103/0019-5154.174011.

38. Wang, X., and Kattan, M.W. (2020). Cohort Studies: Design, Analysis, and Reporting. Chest *158*, S72–S78. https://doi.org/10.1016/j.chest.2020.03.014.

39. Song, J.W., and Chung, K.C. (2010). Observational Studies: Cohort and Case-Control Studies. Plast Reconstr Surg *126*, 2234. https://doi.org/10.1097/PRS.0B013E3181F44ABC.

40. Budach, L., Feuerpfeil, · Moritz, Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, · Felix, and Harmouch, · Hazar (2022). The Effects of Data Quality on Machine Learning Performance. ArXiv.

41. Woodman, R.J., and Mangoni, A.A. (2023). A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future. Aging Clinical and Experimental Research 2023 35:11 *35*, 2363–2397. https://doi.org/10.1007/S40520-023-02552-2.

42. Nguyen, L.H., and Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. PLoS Comput Biol *15*, e1006907. https://doi.org/10.1371/JOURNAL.PCBI.1006907.

43. Harrell, F.E. (2015). Regression Modeling Strategies 2nd ed. (Springer Cham) https://doi.org/10.1007/978-3-319-19425-7.

44. Sorzano, C.O.S., Vargas, J., and Pascual-Montano, A. (2014). A survey of dimensionality reduction techniques. ArXiv.

45. Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., and Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends *1*, 56–70. https://doi.org/10.38094/JASTT1224.

46. Xu, R.F., and Lee, S.J. (2015). Dimensionality reduction by feature clustering for regression problems. Inf Sci (N Y) *299*, 42–57. https://doi.org/10.1016/J.INS.2014.12.003.

47. Moscovich, A., and Rosset, S. (2022). On the Cross-Validation Bias due to Unsupervised Preprocessing. J R Stat Soc Series B Stat Methodol *84*, 1474–1502. https://doi.org/10.1111/RSSB.12537.

48. Bennett, D.A. (2001). How can I deal with missing data in my study? Aust N Z J Public Health *25*, 464–469. https://doi.org/10.1111/J.1467-842X.2001.TB00294.X.

49. Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., and Carpenter, J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ *338*, 157–160. https://doi.org/10.1136/BMJ.B2393.

50. Zhang, Z. (2016). Missing data imputation: focusing on single imputation. Ann Transl Med *4*, 9. https://doi.org/10.3978/J.ISSN.2305-5839.2015.12.38.

51. Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., and Moons, K.G.M. (2006). Review: A gentle introduction to imputation of missing values. J Clin Epidemiol *59*, 1087–1091. https://doi.org/10.1016/J.JCLINEPI.2006.01.014.

52. Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., and Franco, L. (2010). Missing data imputation using statistical and machine learning

methods in a real breast cancer problem. Artif Intell Med *50*, 105–115. https://doi.org/10.1016/J.ARTMED.2010.05.002.

53. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data 2021 8:1 *8*, 1–37. https://doi.org/10.1186/S40537-021-00516-9.

54. Yadav, M.L., and Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. Knowl Based Syst *160*, 104–118. https://doi.org/10.1016/J.KNOSYS.2018.06.012.

55. Dash, C.S.K., Behera, A.K., Dehuri, S., and Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. Decision Analytics Journal *6*, 100164. https://doi.org/10.1016/J.DAJOUR.2023.100164.

56. Kuhn, M., and Johnson, K. (2013). Applied predictive modeling 1st ed. (Springer New York) https://doi.org/10.1007/978-1-4614-6849-3/COVER.

57. Peterson, L.E. (2009). K-nearest neighbor. Scholarpedia *4*, 1883. https://doi.org/10.4249/SCHOLARPEDIA.1883.

58. Singh, D., and Singh, B. (2020). Investigating the impact of data normalization on classification performance. Appl Soft Comput *97*, 105524. https://doi.org/10.1016/J.ASOC.2019.105524.

59. Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., and J. Tallón-Ballesteros, A. (2023). The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. Lecture Notes in Networks and Systems *750 LNNS*, 344–353. https://doi.org/10.1007/978-3-031-42536-3_33.

60. van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., and van der Werf, M.J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics *7*, 142. https://doi.org/10.1186/1471-2164-7-142.

61. Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., and Altman, D.G. (2011). Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms. AJNR Am J Neuroradiol *32*, 437. https://doi.org/10.3174/AJNR.A2425.

62. Altman, D.G. (2014). Categorizing Continuous Variables. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.STAT04857.

63. Bennette, C., and Vickers, A. (2012). Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol *12*, 1–5. https://doi.org/10.1186/1471-2288-12-21/FIGURES/3.

64. Royston, P., Altman, D.G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med *25*, 127–141. https://doi.org/10.1002/SIM.2331.

65. Sauerbrei, W., and Royston, P. (2010). Continuous Variables: To Categorize or to Model? In Proceedings of the Eighth International Conference on Teaching Statistics .

66. Heavner, K.K., Phillips, C. V., Burstyn, I., and Hare, W. (2010). Dichotomization: 2 × 2 (×2 × 2 × 2...) categories: Infinite possibilities. BMC Med Res Methodol *10*, 1–11. https://doi.org/10.1186/1471-2288-10-59/FIGURES/4.

67. Kirk, D., Catal, C., and Tekinerdogan, B. (2022). Predicting Plasma Vitamin C Using Machine Learning. Applied Artificial Intelligence *36*. https://doi.org/10.1080/08839514.2022.2042924.

68. Probst, P., Wright, M.N., and Boulesteix, A.L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov *9*, e1301. https://doi.org/10.1002/WIDM.1301.

69. Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports 2024 14:1 *14*, 1–14. https://doi.org/10.1038/s41598-024-56706-x.

70. Mienye, I.D., and Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. IEEE Access *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287.

71. Breiman, L. (2001). Random forests. Mach Learn *45*, 5–32. https://doi.org/10.1023/A:1010933404324/METRICS.

72. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785.

73. Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. Front Comput Sci *14*, 241–258. https://doi.org/10.1007/S11704-019-8208-Z/METRICS.

74. Bergstra, J., Ca, J.B., and Ca, Y.B. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. Journal of Machine Learning Research *13*, 281–305.

75. Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.L., et al. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdiscip Rev Data Min Knowl Discov *13*, e1484. https://doi.org/10.1002/WIDM.1484.

76. Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering *40*, 16–28. https://doi.org/10.1016/J.COMPELECENG.2013.11.024.

77. Theng, D., and Bhoyar, K.K. (2024). Feature selection techniques for machine learning: a survey of more than two decades of research. Knowl Inf Syst *66*, 1575–1637. https://doi.org/10.1007/S10115-023-02010-5/TABLES/6.

78. Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W., and O'Sullivan, J.M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics *2*. https://doi.org/10.3389/FBINF.2022.927312.

79. Fan, J., and Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. Stat Sin *20*, 101.

80. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., and Liu, H. (2017). Feature Selection. ACM Computing Surveys (CSUR) *50*. https://doi.org/10.1145/3136625.

81. Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics *23*, 2507–2517. https://doi.org/10.1093/BIOINFORMATICS/BTM344.

82. Khaire, U.M., and Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. Journal of King Saud University - Computer and Information Sciences *34*, 1060–1073. https://doi.org/10.1016/J.JKSUCI.2019.06.012.

83. Boehmke, B., and Greenwell, B. (2019). Hands-On Machine Learning with R 1st ed. (Chapman and Hall/CRC) https://doi.org/10.1201/9780367816377.

84. Whalen, S., Schreiber, J., Noble, W.S., and Pollard, K.S. (2021). Navigating the pitfalls of applying machine learning in genomics. Nature Reviews Genetics 2021 23:3 *23*, 169–181. https://doi.org/10.1038/s41576-021-00434-9.

85. Lones, M.A. (2024). How to avoid machine learning pitfalls: a guide for academic researchers.

86. Acharjee, A., Finkers, R., Gf Visser, R., and Maliepaard, C. (2013). Comparison of Regularized Regression Methods for ~Omics Data. Metabolomics *3*, 126. https://doi.org/10.4172/2153-0769.1000126.

87. Acharjee, A., Larkman, J., Xu, Y., Cardoso, V.R., and Gkoutos, G. V. (2020). A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC Med Genomics *13*, 1–14. https://doi.org/10.1186/S12920-020-00826-6/FIGURES/5.

88. Dinga, R., Penninx, B.W.J.H., Veltman, D.J., Schmaal, L., and Marquand, A.F. (2019). Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. bioRxiv, 743138. https://doi.org/10.1101/743138.

89. Naser, M.Z., and Alavi, A. (2020). Insights into Performance Fitness and Error Metrics for Machine Learning. Architecture, Structures and Construction *3*, 499–517. https://doi.org/10.1007/s44150-021-00015-8.

90. Chicco, D., Warrens, M.J., and Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623/SUPP-1.

91. Palacio-Niño, J.-O., and Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms.

92. Raschka, Sebastian., and Mirjalili, Vahid. (2019). Python machine learning : machine learning and deep learning with python, scikit-learn, and tensorflow 2 3rd ed. (Packt Publishing, Limited).

93. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., and Johannes, R.S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261.

94. Wainer, J., and Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst Appl *182*, 115222. https://doi.org/10.1016/J.ESWA.2021.115222.

95. Cawley, G.C., and Talbot, N.L.C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research *11*, 2079–2107.

96. Steyerberg, E.W., and Harrell, F.E. (2016). Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol *69*, 245–247. https://doi.org/10.1016/J.JCLINEPI.2015.04.005.

97. Van Calster, B., McLernon, D.J., Van Smeden, M., Wynants, L., Steyerberg, E.W., Bossuyt, P., Collins, G.S., MacAskill, P., Moons, K.G.M., and Vickers, A.J. (2019). Calibration: The Achilles heel of predictive analytics. BMC Med *17*, 1–7. https://doi.org/10.1186/S12916-019-1466-7/TABLES/1.

98. Doshi-Velez, F., and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning.

99. Dormann, C.F. (2020). Calibration of probability predictions from machine-learning and statistical models. Global Ecology and Biogeography *29*, 760–765. https://doi.org/10.1111/GEB.13070.

100. Kull, M., De Menezes, T., Filho, S., and Flach, P. (2017). Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (PMLR), pp. 623–631.

101. Niculescu-Mizil, A., and Caruana, R. (2005). Predicting good probabilities with supervised learning. ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning, 625–632. https://doi.org/10.1145/1102351.1102430.

102. Platt, J.C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv. Large Margin Classif *10*, 61–74.

103. Ho, S.Y., Phua, K., Wong, L., and Bin Goh, W.W. (2020). Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns *1*, 100129. https://doi.org/10.1016/J.PATTER.2020.100129.

104. de Hond, A.A.H., Leeuwenberg, A.M., Hooft, L., Kant, I.M.J., Nijman, S.W.J., van Os, H.J.A., Aardoom, J.J., Debray, T.P.A., Schuit, E., van Smeden, M., et al. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med *5*. https://doi.org/10.1038/S41746-021-00549-7.

105. Cabitza, F., Campagner, A., Soares, F., García de Guadiana-Romualdo, L., Challa, F., Sulejmani, A., Seghezzi, M., and Carobene, A. (2021). The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed *208*, 106288. https://doi.org/10.1016/J.CMPB.2021.106288.

106. Ramspek, C.L., Jager, K.J., Dekker, F.W., Zoccali, C., and Van DIepen, M. (2021). External validation of prognostic models: what, why, how, when and where? Clin Kidney J *14*, 49–58. https://doi.org/10.1093/CKJ/SFAA188.

107. Siontis, G.C.M., Tzoulaki, I., Castaldi, P.J., and Ioannidis, J.P.A. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol *68*, 25–34. https://doi.org/10.1016/J.JCLINEPI.2014.09.007.

108. Heshiki, Y., Vazquez-Uribe, R., Li, J., Ni, Y., Quainoo, S., Imamovic, L., Li, J., Sørensen, M., Chow, B.K.C., Weiss, G.J., et al. (2020). Predictable modulation of cancer treatment outcomes by the gut microbiota. Microbiome *8*. https://doi.org/10.1186/S40168-020-00811-2.

109. Van Calster, B., Steyerberg, E.W., Wynants, L., and van Smeden, M. (2023). There is no such thing as a validated prediction model. BMC Med *21*, 1–8. https://doi.org/10.1186/S12916-023-02779-W/FIGURES/2.

110. la Roi-Teeuw, H.M., van Royen, F.S., de Hond, A., Zahra, A., de Vries, S., Bartels, R., Carriero, A.J., van Doorn, S., Dunias, Z.S., Kant, I., et al. (2024). Don't be misled: 3 misconceptions about external validation of clinical prediction models. J Clin Epidemiol *172*. https://doi.org/10.1016/j.jclinepi.2024.111387.

111. Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 2021, Vol. 23, Page 18 *23*, 18. https://doi.org/10.3390/E23010018.

112. Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) *13200 LNAI,* 39–68. https://doi.org/10.1007/978-3-031-04083-2_4/FIGURES/7.

113. Lundberg, S.M., Allen, P.G., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In 31st Conference on Neural Information Processing Systems https://doi.org/https://doi.org/10.48550/arXiv.1705.07874.

114. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining *13-17-August-2016*, 1135–1144. https://doi.org/10.1145/2939672.2939778/SUPPL_FILE/KDD2016_RIBEIRO_ANY_CLASSIFIER_01-ACM.MP4.

115. Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics *26*, 1340–1347. https://doi.org/10.1093/BIOINFORMATICS/BTQ134.

116. Strobl, C., Boulesteix, A.L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics *8*, 1–21. https://doi.org/10.1186/1471-2105-8-25/FIGURES/11.

117. Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. BMC Bioinformatics *9*, 1–11. https://doi.org/10.1186/1471-2105-9-307/FIGURES/4.

118. Toloşi, L., and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics *27*, 1986–1994. https://doi.org/10.1093/BIOINFORMATICS/BTR300.

119. Saarela, M., and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Appl Sci *3*, 1–12. https://doi.org/10.1007/S42452-021-04148-9/TABLES/4.

120. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A *116*, 22071–22080. https://doi.org/10.1073/PNAS.1900654116/SUPPL_FILE/PNAS.1900654116.SAPP.PDF.

121. Wang, H., Yang, F., and Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. BMC Bioinformatics *17*, 1–18. https://doi.org/10.1186/S12859-016-0900-5/FIGURES/9.

122. Kalousis, A., Prados, J., and Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. Knowl Inf Syst *12*, 95–116. https://doi.org/10.1007/S10115-006-0040-8/METRICS.

123. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell *1*, 206. https://doi.org/10.1038/S42256-019-0048-X.

124. de Bruijn, H., Warnier, M., and Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Gov Inf Q *39*, 101666. https://doi.org/10.1016/J.GIQ.2021.101666.

125. Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, · Xiang, Li, R., Yao, N., Wang, · Xinyi, Gu, · Xiaotong, et al. (2023). Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. Human-Centric Intelligent Systems 2023 3:3 *3*, 161–188. https://doi.org/10.1007/S44230-023-00038-Y.

126. Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. PLoS Med *2*, 2–8. https://doi.org/10.1371/JOURNAL.PMED.0020124.

127. Ioannidis, J.P.A. (2014). How to Make More Published Research True. PLoS Med *11*, e1001747. https://doi.org/10.1371/JOURNAL.PMED.1001747.

128. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature *533*, 452–454. https://doi.org/10.1038/533452A.

129. Begley, C.G., and Ioannidis, J.P.A. (2015). Reproducibility in Science. Circ Res *116*, 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819.

130. McDermott, M.B.A., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., and Ghassemi, M. (2021). Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med *13*. https://doi.org/10.1126/SCITRANSLMED.ABB1655.

131. Beam, A.L., Manrai, A.K., and Ghassemi, M. (2020). Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA *323*, 305–306. https://doi.org/10.1001/JAMA.2019.20866.

132. Heil, B.J., Hoffman, M.M., Markowetz, F., Lee, S.I., Greene, C.S., and Hicks, S.C. (2021). Reproducibility standards for machine learning in the life sciences. Nat Methods *18*, 1132–1135. https://doi.org/10.1038/S41592-021-01256-7.

133. Stefan, A.M., and Schönbrodt, F.D. (2023). Big little lies: a compendium and simulation of p-hacking strategies. R Soc Open Sci *10*. https://doi.org/10.1098/RSOS.220346.

Figures

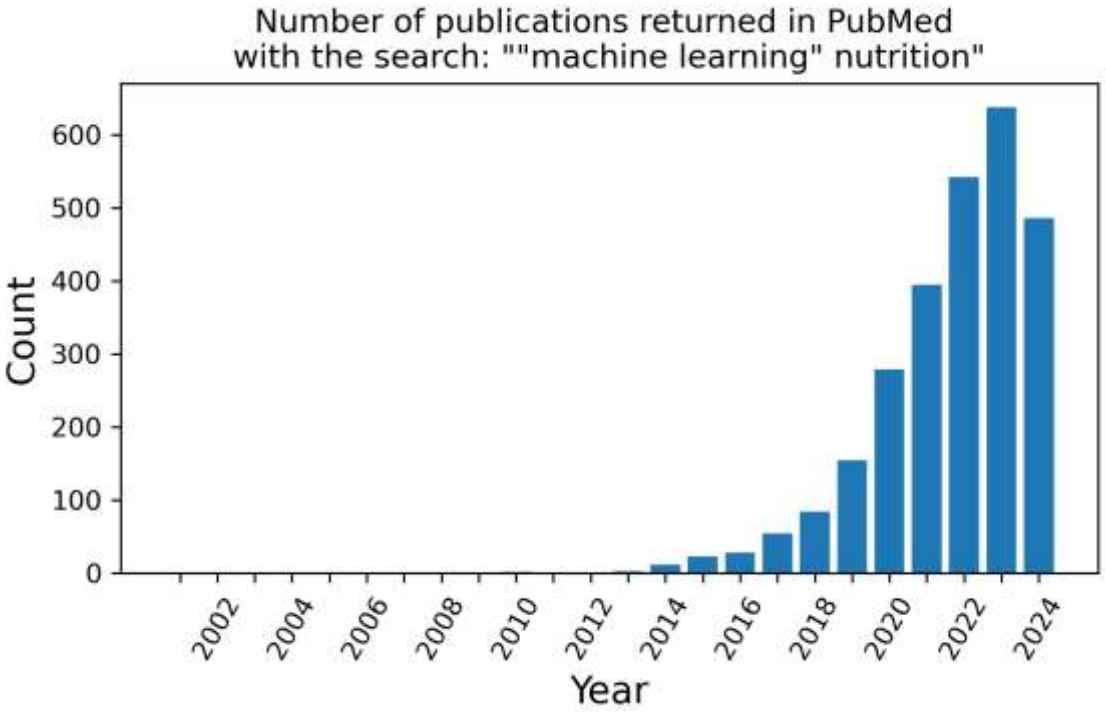

Figure 1: The number of publications by year returned in PubMed with the search terms ""machine learning" nutrition" from 2001 (the first year containing a publication with these search terms) to 2024. Date of search: 20th August 2024, 11:13 BST.

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

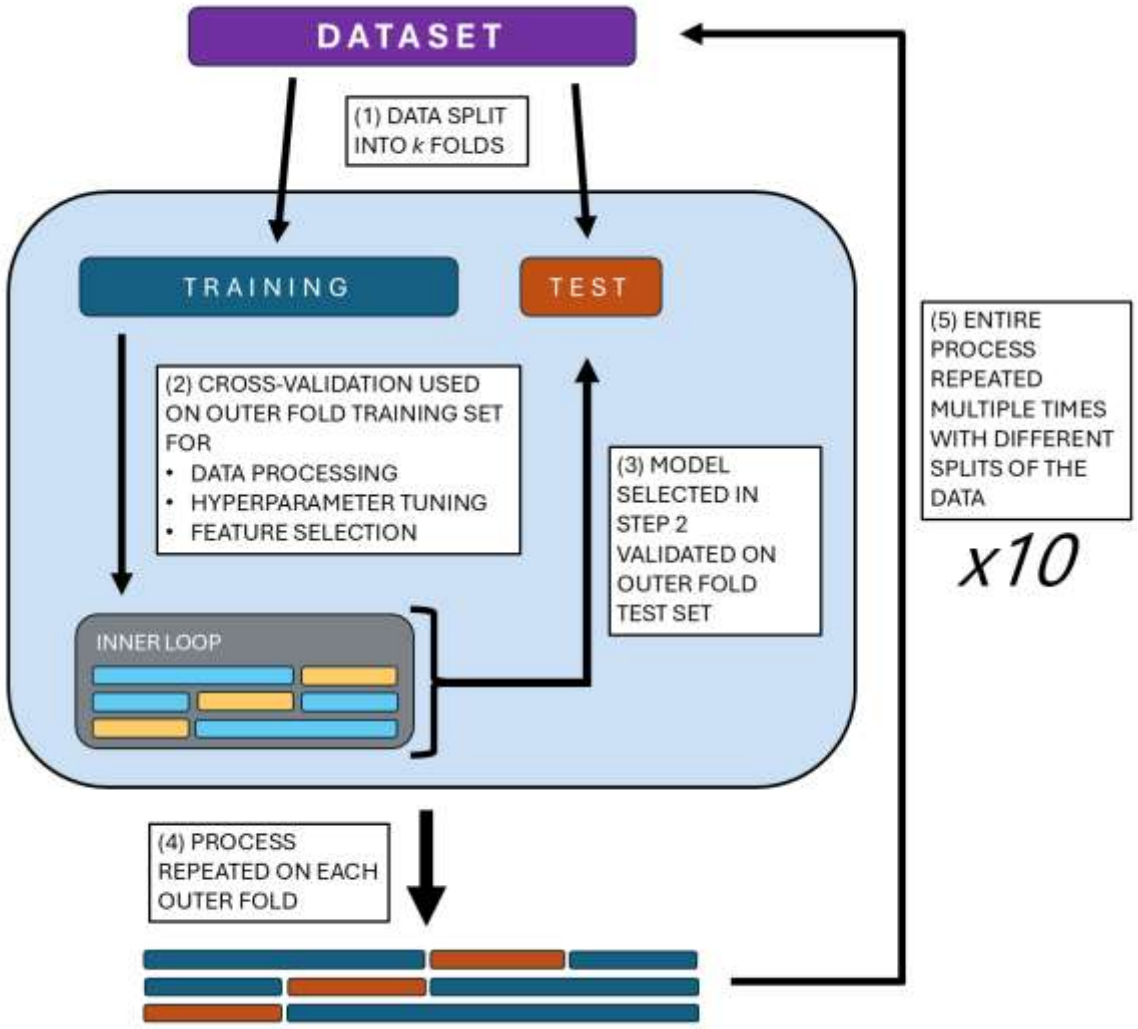

Figure 2: A robust internal validation scheme using nested cross-validation. Data are first split using cross-validation (outer loop; step 1). In each fold of the outer loop, cross-validation is used on the training data (dark blue) for data processing, hyperparameter optimisation and feature selection. This is known as the inner loop (grey box; step 2). The performance of the model selected in the inner loop is then validated on the outer fold test data (dark orange; step 3). This process is depicted in the large, light blue box, and is repeated in each fold of the outer loop (step 4). The whole process is then repeated multiple times to account for instability of the results depending on how the dataset is split (step 5).

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

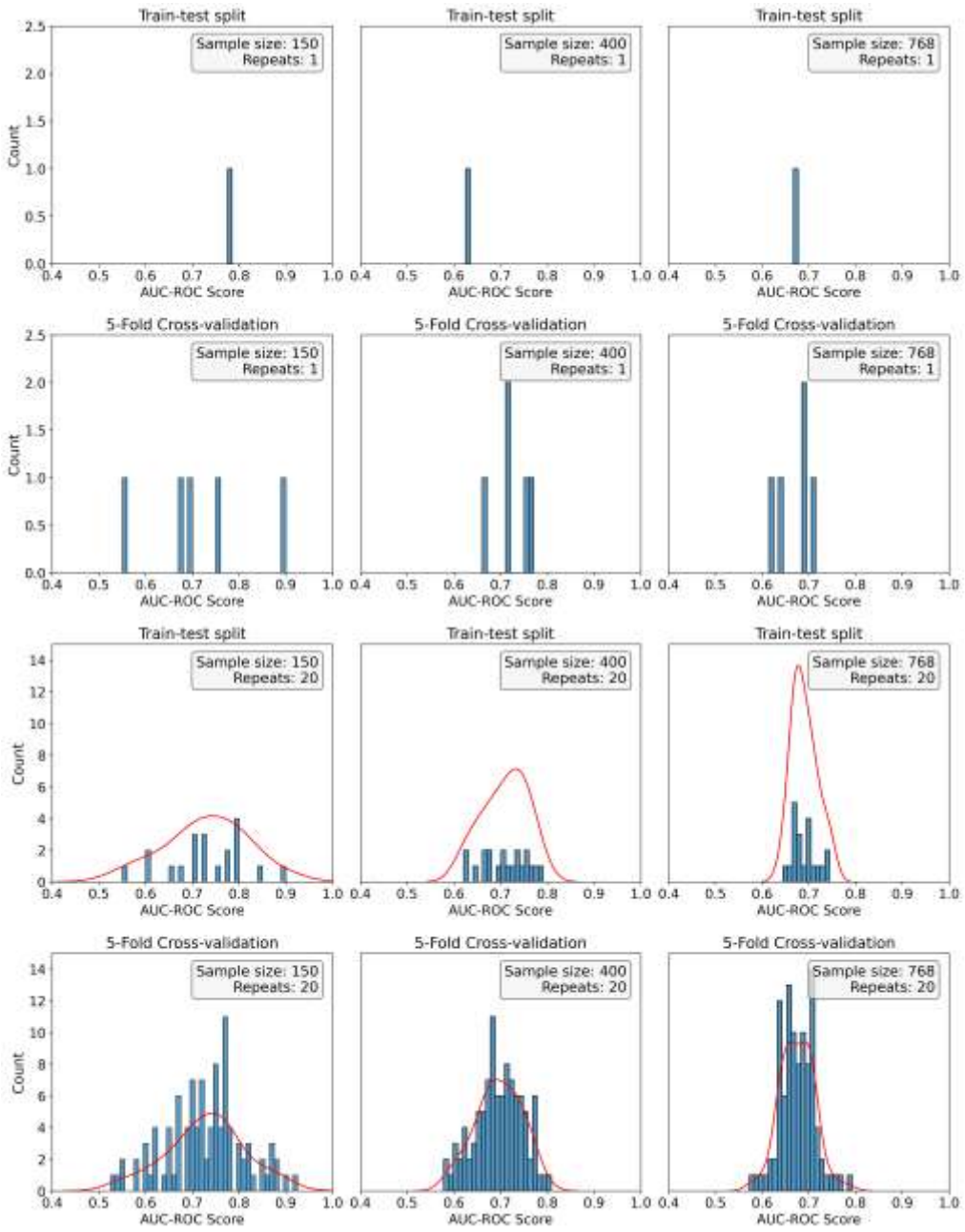

Figure 3: The effect of validation technique, the number of times it is repeated, and sample size on the stability and uncertainty of the results. One repeat of cross-validation (second row) is an immediate improvement over one repeat of train-test split (first row) because the perturbation in the training and test data in each fold provide an indication of the stability of the AUC-ROC scores. Repeating the validation procedure multiple times with different subsamples also allows stability to be estimated, with this being more effective in cross-validation (fourth row) than train-test split (third row) because there are more test scores. Both instability and uncertainty tend to decrease as sample size increases.

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

Table 1: A summary of key points and common pitfalls in each step in the machine learning procedure for research.

Step	Key points	Common pitfalls
1. Data Collection	- Data quality should be a priority - Sample should be representative of target population and adequately sized - Data longevity, questionnaire wording, variables collected, data storage, ethical considerations, and documentation and metadata are important when constructing a dataset	- Variables potentially relevant to the problem are not collected - Sample size is too small to provide reliable results - Variables become unusable or irrelevant (i.e., questionnaire questions) over time
2. Data Processing	- Data processing steps should be meticulously documented - Data processing steps that make use of the outcome should be performed within validation steps	- Including or excluding features based on their relationship with the outcome variable - Imputing missing values with the mean, median or mode - Defining outliers based on general rules without regard for the specific characteristics of the data - Discretising continuous variables
3. Modelling	- Different algorithms can be used, and their results compared - Predictive performance, speed and interpretability are key factors in determining algorithms to be considered - Hyperparameters should be tuned to optimise performance	- Optimising the hyperparameters of some but not all algorithms being compared - Overstating the relevance of small differences in predictive capacity between models - Information leakage during data processing, hyperparameter tuning

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

	- Robust internal validation schemes should be used to improve the reliability of the results	and feature selection
4. Evaluation	- Metrics should be chosen based on their relevance to the problem and the intended application of the model - Different metrics evaluate performance from different perspectives - Clustering can be repeated with different metrics and the results compared - Validation procedures should be repeated multiple times to account for instability in the results	- The metrics used are inappropriate, or provide an incomplete or biased evaluation of model performance - Validation procedures are applied only once and not repeated, meaning the stability of the results cannot be known - Validation techniques do not account for imbalanced data or dependent observations
5. Interpretation	- Evaluation metrics must be properly understood to allow proper interpretation of the results - Multiple test scores may be described with summary statistics or presented on plots - Feature importance can be estimated in multiple ways; there is no single best approach	- Metrics are misinterpreted - One single score from cross-validation is reported; how individual test scores were aggregated is not described - Default feature importance methods are used - Explainability AI techniques that are not suitable for the dataset are used - Multicollinearity is not accounted for during feature importance estimation - Feature importance estimates from poorly fit models are reported

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press

		- Feature importance estimates are derived from one split of the data; feature importance instability is not accounted for
6. Reporting	- All steps, from obtaining the data to reporting the methodology and results, should be described completely and transparently - If possible, code should be published	- Parts of the data processing, modelling or evaluation are missing or incompletely described - Code is unavailable or difficult to read - Only positive findings are reported, and those that make the results seem less convincing are omitted

https://doi.org/10.1017/S0029665124007638 Published online by Cambridge University Press