

8-22-2023

## How I Read an Article That Uses Machine Learning Methods

Aziz Nazha

Thomas Jefferson University, [aziz.nazha@jefferson.edu](mailto:aziz.nazha@jefferson.edu)


Olivier Elemento

Shannon McWeeney

Moses Miles

Torsten Haferlach

Follow this and additional works at: <https://jdc.jefferson.edu/kimmelccfp>

 Part of the [Other Communication Commons](#), [Reading and Language Commons](#), and the [Theory and Algorithms Commons](#)

[Let us know how access to this document benefits you](#)

---

### Recommended Citation

Nazha, Aziz; Elemento, Olivier; McWeeney, Shannon; Miles, Moses; and Haferlach, Torsten, "How I Read an Article That Uses Machine Learning Methods" (2023). *Kimmel Cancer Center Faculty Papers*. Paper 108. <https://jdc.jefferson.edu/kimmelccfp/108>

This Article is brought to you for free and open access by the Jefferson Digital Commons. The Jefferson Digital Commons is a service of Thomas Jefferson University's [Center for Teaching and Learning \(CTL\)](#). The Commons is a showcase for Jefferson books and journals, peer-reviewed scholarly publications, unique historical collections from the University archives, and teaching tools. The Jefferson Digital Commons allows researchers and interested readers anywhere in the world to learn about and keep up to date with Jefferson scholarship. This article has been accepted for inclusion in Kimmel Cancer Center Faculty Papers by an authorized administrator of the Jefferson Digital Commons. For more information, please contact: [JeffersonDigitalCommons@jefferson.edu](mailto:JeffersonDigitalCommons@jefferson.edu).

## TO THE EDITOR:

## How I read an article that uses machine learning methods

Aziz Nazha,<sup>1</sup> Olivier Elemento,<sup>2</sup> Shannon McWeeney,<sup>3</sup> Moses Miles,<sup>4</sup> and Torsten Haferlach,<sup>5</sup> on behalf of the American Society of Hematology Artificial Intelligence Taskforce

<sup>1</sup>Department of Oncology, Sidney Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, PA; <sup>2</sup>Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York City, NY; <sup>3</sup>Division of Bioinformatics and Computational Biology, Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR; <sup>4</sup>American Thrombosis and Hemostasis Network, Rochester, NY; and <sup>5</sup>Munich Leukemia Laboratory, Munich, Germany

Machine learning (ML) has revolutionized many industries including the health care industry by providing innovative solutions to some of the most pressing problems. With the advancement of technology and increasing amounts of data being generated, ML has become a central tool for health care professionals in various fields, such as diagnostics, drug discovery, and personalized medicine.<sup>1-6</sup> The ability of ML algorithms to analyze vast amounts of complex data has led to improved accuracy and speed in diagnosis, better targeting of treatments, and more personalized care for patients.

Reading a scientific paper that uses ML methodologies can be a challenging task for those who are not familiar with the field.<sup>6</sup> However, with a clear understanding of the basic concepts and a critical approach, it is possible to gain valuable insights from these papers. In this commentary, we will provide a step-by-step guide on how to read a scientific paper that has ML methodologies.

**Step 1: Understand the problem being addressed.** The first step in reading an ML paper is to understand the problem that the authors are trying to solve and, more importantly, understand the clinical or scientific impact of solving this problem.<sup>7</sup> In other words, if the aim of the study is to solve a clinical problem, how does the answer or the recommendation provided by the algorithm help physicians or researchers in their day-to-day practice, and is this solution mature enough to be implemented in clinical workflows? Major clinical problems in health care can mainly affect either patient outcomes or operations (can I make the process easier and faster for the patient and the health care system?).

**Step 2: Assess the quality of the data.** The quality of the data used to build the ML model is crucial for the validity of the results. Following are some questions that can be used to evaluate the data:

1. **Sample size:** Is the size of the training, validation, and test sets enough to build a reproducible and generalizable ML model? Is this size of the data appropriate for the chosen methods (ie, some methods are “data-hungry” and understanding which methods require larger datasets is key)? However, different algorithms require different data types (image, tabular, text, or others) and sizes, and there are no rules of thumb or formulas that can estimate the perfect data.
2. **Relevance:** Are the data appropriate and relevant to the problem that the model is trying to solve?
3. **Accuracy:** How are the data collected and annotated (human vs natural language process). How are the data transformed to make it ready for ML use, etc.
4. **Consistency:** Are the data consistent? Do they have any missing values and how the authors dealt with this?
5. **Representativeness:** The data should be representative of the population being studied.
6. **Balance:** The data should be balanced, with roughly equal representation of all relevant classes or groups. However, most health care data are unbalanced. It is critical to understand how the authors dealt with unbalanced data.

**Table 1. ML terminologies**

Term	Description
AI	AI refers to the simulation of human intelligence in machines that are programmed to think and act like humans. These machines use algorithms and data to perform tasks, such as recognizing patterns, learning from experience, making decisions, and solving problems.
ML	ML is a subfield of AI that involves training algorithms to make predictions or decisions based on data. ML algorithms use statistical models and algorithms to analyze data, learn from that data, and then make a prediction or classification about new data. The goal of ML is to automatically improve the performance of the algorithm over time by learning from the input data. This allows the algorithm to become more accurate in its predictions and decisions as it is exposed to more data.
Deep learning	Deep learning is a subset of ML that uses artificial neural networks with multiple layers to process and analyze complex data. These deep neural networks are trained to identify patterns in large amounts of data, such as images, speech, and text, and then make predictions or decisions based on that data. Deep learning has led to breakthroughs in many applications of AI, including computer vision, speech recognition, and natural language processing.
Transfer learning	Transfer learning is an ML technique where a model trained on one task is reused as the starting point for a model on a related task. The idea is that the knowledge gained from solving one problem can be useful for solving a similar problem, allowing for faster training times and improved performance compared with training a model from scratch. Transfer learning is commonly used in computer vision and natural language processing, where models trained on large, general-purpose data sets can be fine-tuned for specific tasks with smaller amounts of data.
Supervised learning	Supervised learning is a type of ML technique in which an algorithm learns to make predictions or decisions by training on labeled data. The goal of supervised learning is to learn a mapping function that can accurately predict output values for new input data. The learning process involves adjusting the parameters of the algorithm based on the errors between the predicted and actual output values. The most common types of supervised learning algorithms are regression and classification. Common types of supervised learning algorithms are decision trees, linear regression, logistic regression, random forest, support vector machine, and others.
Unsupervised learning	Unsupervised learning is a type of ML technique in which an algorithm learns to identify patterns and relationships in data without being explicitly trained on labeled data. In unsupervised learning, the algorithm is presented with a set of input data, and it learns to discover patterns and structure within the data on its own. Some common types of unsupervised learning algorithms are clustering, dimensionality reduction, and anomaly detection.
Semisupervised learning	Semisupervised learning is a type of ML technique that combines both labeled and unlabeled data for training a model. In semisupervised learning, the algorithm is presented with a small amount of labeled data and a large amount of unlabeled data, and it learns to make predictions by using both types of data.
CNN	CNN is a type of deep learning artificial neural network used for image and video recognition, as well as natural language processing tasks. It is designed to process data through multiple layers of arrays, called convolutions, which learn features from the input data, reducing its dimensionality and allowing for pattern recognition. The use of pooling layers and fully connected layers allows a CNN to make predictions based on the features it has learned, making it a powerful tool in computer vision and NLP.
RNN	RNNs are a type of artificial neural network used for processing sequential data such as time series, natural language text, and speech. Unlike traditional feedforward neural networks, RNNs have a feedback loop that allows information to persist, allowing the network to maintain information from past inputs and use it in conjunction with current inputs in making predictions. This makes RNNs well suited for tasks in which the current output is dependent on the previous inputs, such as language generation, speech recognition, and machine translation.
NLP	NLP is a field of AI concerned with enabling computers to understand, interpret, and generate human language. NLP encompasses a wide range of tasks, including text classification, sentiment analysis, machine translation, named entity recognition, and question answering, among others. To perform NLP tasks, techniques from computational linguistics, computer science, and ML are combined to develop algorithms that can process and analyze large amounts of text and speech data. The goal of NLP is to enable computers to understand and process human language in a way that is similar to how humans do, making it a key component in the development of artificial general intelligence.
GNNs	GNNs are a type of neural network used for processing data structured as graphs, such as social networks, molecule structures, and road networks. GNNs are designed to handle the non-Euclidean structure of graph data by updating node representations based on the representations of their neighboring nodes, and by propagating information through edges in the graph. This allows GNNs to learn rich representations of the graph structure, making them useful for tasks such as node classification, graph classification, and link prediction. GNNs have been shown to outperform traditional neural networks on many graph-based tasks and have become a rapidly growing area of research in the field of deep learning.
Algorithm	An algorithm is a set of instructions or rules that a computer follows to perform a specific task. In ML, algorithms are used to analyze data and make predictions based on that data.
Model	A model is a representation of a system or process that can be used to make predictions. In ML, models are trained on data and then used to make predictions about new data.
Parameter	A parameter is a configuration variable that is internal to the model and whose value is learned from the data during training. It is a part of the model that can be adjusted to optimize the performance of the algorithm. For example, in linear regression, the parameters are the coefficients of the model that are learned during training to minimize the error between the predicted values and the actual values. In neural networks, the parameters include the weights and biases that are learned through backpropagation during training.
Hyperparameter	A hyperparameter is a setting or configuration that is external to the model and is used to control its learning process. Unlike model parameters that are learned during training, hyperparameters are set before the learning process begins and remain constant throughout training. Examples of hyperparameters include the learning rate of the model, the number of hidden layers in a neural network, the number of decision trees in a random forest model, and regularization parameters. The selection of appropriate hyperparameters can have a significant impact on the performance of the model.
Training data	Training data is a set of data that is used to train an ML model. The model uses the training data to learn how to make predictions. In this data set, the features and outcomes are known, and therefore, the model can learn how to predict the outcomes from the features.
Validation data	Validation data are a set of data that are used to evaluate the performance of an ML model. The model is tested on the validation data to see how well it makes predictions.

AI, artificial intelligence; AUC, area under the curve; CNN, convolutional neural network; GNN, graph neural network; NLP, natural language processing; RNN, recurrent neural network; ROC, receiver operating characteristic.

**Table 1 (continued)**

Term	Description
Test data	Test data are a set of data that provide a final, real-world check of an unseen data set to confirm that the ML algorithm was trained effectively. Preferably, the test data set should represent an external, multi-institutional data set. Furthermore, the outcomes are known in this data set but are not used in training and therefore can be used to test the model.
Overfitting	Overfitting occurs when an ML model is too closely fit to the training data, resulting in poor performance on new data.
Underfitting	Underfitting occurs when an ML model is too simple to capture the complexity of the data, resulting in poor performance on both the training and validation data.
Bias	Bias occurs when an ML model makes systematic errors in its predictions. Bias can result from errors in the data or from a poorly designed model.
Accuracy	Accuracy is a measure of how well an ML model makes predictions compared with the actual value.
ROC/AUC	ROC is a commonly used evaluation metric in binary classification problems. It plots the true-positive rate against the false-positive rate at various thresholds, providing a visual representation of the classifier's performance. AUC is a single number summary of the ROC curve, representing the overall performance of a classifier. It measures the area under the ROC curve and ranges from 0 to 1, with a higher AUC indicating a better classifier. An AUC of 0.5 indicates a classifier with no discrimination power, whereas an AUC of 1 means perfect discrimination.
Precision	Precision is a measure of how many of the predictions made by an ML model are correct.
Recall	Recall is a measure of how many of the actual values are correctly predicted by an ML model.
F1 score	F1 score is a commonly used metric to evaluate the performance of binary and multiclass classification algorithms. It is the harmonic mean of precision and recall, in which precision is the number of true-positive predictions divided by the sum of true-positive and false-positive predictions, and recall is the number of true-positive predictions divided by the sum of true-positive and false-negative predictions. The F1 score ranges from 0 to 1, with a higher score indicating better performance. It balances precision and recall, making it a useful metric when the cost of false negatives and false positives is not equal.

AI, artificial intelligence; AUC, area under the curve; CNN, convolutional neural network; GNN, graph neural network; NLP, natural language processing; RNN, recurrent neural network; ROC, receiver operating characteristic.

7. Bias: To evaluate bias in data, it is important to look at the distribution of certain characteristics, such as race, gender, or socioeconomic status, among the samples in the data set,<sup>8</sup> and how the data were collected. This will help to identify any disparities or overrepresentation of certain groups, which can indicate the presence of bias in the data. It is critical to evaluate bias at this stage because if this is not addressed properly, it could produce a biased model.<sup>8-10</sup>

**Step 3: Familiarize yourself with the ML methods used.** The next step is to understand the ML methods that the authors have used to solve the problem. Many papers will provide a brief overview of the methods used (in clinical or applied journals), but it is important to have a good understanding of the underlying concepts.<sup>1-6</sup> It is critical to familiarize yourself with some of these terminologies presented in Table 1. There are many papers that explain these terminologies in a very simple manner.<sup>1-6</sup> It is also important to understand the key issues in building ML (Figure 1) models and what the authors did to address these at each step.

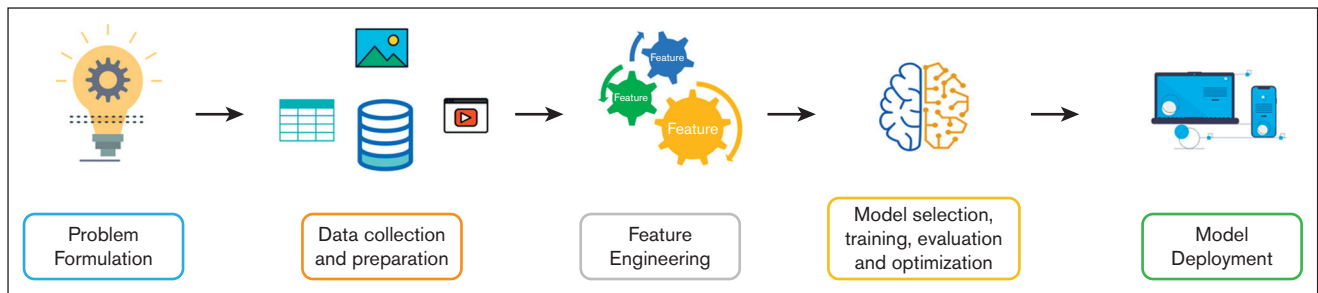
**Step 4: Evaluate the results and how they are presented in the paper.** Ask these questions:

1. How did authors divide the cohort (eg, training, validation, and test)? Ideally, the test cohort should be completely different from the original one. From where and how these cohorts are collected? Is this a single-center study or a study from multiple centers? Are these centers present in 1 country or worldwide? Is there any testing of the model at another site?
2. How did the authors report the efficacy of the model? Reporting the accuracy and area under the curve (AUC) can be misleading especially when the data are unbalanced (eg, if the authors are building ML to predict bleeding in the brain on computed tomography that occurs in 3% of the scans, a model

can be correct 97% of the time by answering no bleed, but this model is not useful clinically). Authors should report the entire confusion matrix (false positive, true positive, false negative, and true negative) among other important matrices, such as precision, recall, precision or recall AUC, and F1 score, and other matrices as deemed important to the type of ML algorithm applied (Table 1). It is important to report these matrices on the evaluation and test cohorts not on the training cohorts.

3. Is there evidence of overfitting or underfitting (Table 1)? To evaluate evidence of overfitting or underfitting in ML, one can examine the training and validation accuracy (if the training accuracy is much higher than the validation accuracy, it could be a sign of overfitting, and if it is lower, it could be a sign of underfitting), learning curves, cross-validation results, and test set performance. These techniques provide insight into whether the model is overfitting or underfitting the data and can help in selecting an appropriate model with optimal performance.
4. Is the model (and its subsequent predictions) explainable? Explainability of the ML models in health care is very important. This will allow the end user (health care provider, researcher, etc) to understand the model and learn from it but, more importantly, assure that the model is not using patterns in the image or data set that are irrelevant to make the final prediction. Several studies have shown that some deep learning algorithm that evaluates outcomes in imaging data can detect areas not of interest on the image.<sup>11</sup> Some ML models could be useful without explainability, if constructed and validated properly.

**Step 5: Critically evaluate the conclusions and implications.** Finally, it is important to critically evaluate the conclusions and implications of the study and whether the result support the conclusion. This includes considering the limitations of the study,



**Figure 1. Steps to build a machine learning model.** Problem formulation: The first step is to clearly define the problem that you want to solve. This involves defining the inputs and outputs of your model, as well as the type of problem you are trying to solve (classification, regression, clustering, etc). It is important to have a clear understanding of the problem you are trying to solve before you start building a model. Data collection: Once you have formulated the problem, the next step is to collect the relevant data. This may involve scraping data from websites, downloading data sets from public repositories, or collecting data through surveys or experiments. It is important to collect enough data to train your model and validate its performance. Data preparation: After collecting the data, you will need to clean and preprocess it. This involves removing any irrelevant data, dealing with missing values, and transforming the data into a suitable format for ML algorithms. It also includes dividing the data set into training, validation, and test cohorts. This step can take a lot of time and effort, but it is essential for building an accurate and effective model. Feature engineering: Feature engineering is the process of selecting and transforming the input variables (features) in a way that will improve the performance of the model. This may involve selecting the most relevant features, transforming them into a different representation (eg, using one-hot encoding), or creating new features based on existing ones. Feature engineering can have a significant impact on the performance of the model. Model selection: Once you have prepared the data and engineered the features, the next step is to select a suitable ML algorithm. This involves choosing the type of algorithm (eg, decision trees, neural networks, support vector machines) and the specific parameters of the algorithm. This step requires some knowledge of ML and experience with different algorithms. Model training: After selecting the algorithm, the next step is to train the model on the prepared data. This involves feeding the input data into the algorithm and adjusting the model parameters to optimize its performance. This step can take a lot of time and computational resources, especially for large data sets and complex models. Model evaluation: Once the model has been trained, the next step is to evaluate its performance on a separate test set of data. This involves measuring metrics, such as accuracy, precision, recall, and F1 score, to assess the performance of the model. It is important to test the model on data that it has not seen before to ensure that it can be generalized to new data. Model optimization: If the model performance is not satisfactory, then the next step is to optimize the model. This involves tweaking the model parameters, changing the algorithm, or modifying the feature engineering process to improve the model's performance. This step may require several iterations until the desired level of performance is achieved. Model deployment: Once you have built a satisfactory model, the final step is to deploy it in a production environment. This may involve integrating the model into a web application, creating an application programming interface for other developers to use, or deploying it as a stand-alone application. It is important to ensure that the model is well documented and tested thoroughly before it is deployed.

generalizability of the results, and potential impact of the findings on the field. More importantly, it is important to consider the practical deployment of the ML model developed in the study if the study intention is to develop a novel model rather than using ML as an analytic tool. Deployment options for ML models include developing a user-friendly interface for inputting data and receiving outputs, integrating the model into an electronic health care record or imaging database within a hospital, or other methods. Regardless of the chosen deployment strategy, it is essential for the authors to outline their plans for making the model accessible to the public and to address the steps they will take to deploy the model after publication.

With the widespread use of ML methodology in scientific papers, it has become important for all physicians and researchers to comprehend the processes of building, validating, and deploying these models. This will enable us to distinguish between poor scientific studies, comprehend the strengths and limitations of these algorithms, and learn how to overcome them.

**Contribution:** A.N. wrote the initial draft, and O.E., S.M., T.H., and M.M. reviewed, edited, and approved the final manuscript.

**Conflict-of-interest disclosure:** A.N. works at Incyte Pharma and owns stocks at Incyte and Amazon. T.H. works at Munich Leukemia Laboratory. The remaining authors declare no competing financial interests.

See "Appendix" for members of the American Society of Hematology Artificial Intelligence Taskforce.

**Correspondence:** Aziz Nazha, Thomas Jefferson University, 1007 Stewart St, Philadelphia, PA 98101; email: [ANazha@incyte.com](mailto:ANazha@incyte.com).

**Appendix:** The members of the American Society of Hematology Artificial Intelligence Taskforce are Aziz Nazha, Olivier Elemento, Shannon McWeeney, Moses Miles, and Torsten Haferlach.

## References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.
3. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019;18(6):463-477.
4. Radakovich N, Nagy M, Nazha A. Machine learning in haematological malignancies. *Lancet Haematol*. 2020;7(7):e541-e550.
5. Nagy M, Radakovich N, Nazha A. Machine learning in oncology: what should clinicians know? *JCO Clin Cancer Inform*. 2020;4:799-810.
6. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019;322(18):1806-1816.

7. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.
8. Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. *Commun Med.* 2021;1(1):25.
9. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med.* 2021;27(12):2176-2182.
10. Ravi N, Chaturvedi P, Huerta EA, et al. FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. *Sci Data.* 2022;9(1):657.
11. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell.* 2021;3(7):610-619.