PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Eight quick tips for biologically and medically informed machine learning

Luca Oneto[1], **Davide Chicco**[2,3]*

1 Dipartimento di Informatica Bioingegneria Robotica e Ingegneria dei Sistemi, Università di Genova, Genoa, Italy, **2** Dipartimento di Informatica Sistemistica e Comunicazione, Università di Milano-Bicocca, Milan, Italy, **3** Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

* davidechicco@davidechicco.it

OPEN ACCESS

Citation: Oneto L, Chicco D (2025) Eight quick tips for biologically and medically informed machine learning. PLoS Comput Biol 21(1): e1012711. https://doi.org/10.1371/journal.pcbi.1012711

Editor: Patricia M. Palagi, SIB Swiss Institute of Bioinformatics, SWITZERLAND

Published: January 9, 2025

Funding: The work of L.O. is partially supported by FAIR (PE00000013) project under the National Recovery and Resilience Plan of Ministero dell'Università e della Ricerca of Italy program funded by the the European Union – Next Generation EU programme. The work of D.C. is funded by the European Union – Next Generation EU programme, in the context of the National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 "Conseguenze e sfide dell'invecchiamento", Project Age-It (Ageing Well in an Ageing Society) and partially supported by Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no

Abstract

Machine learning has become a powerful tool for computational analysis in the biomedical sciences, with its effectiveness significantly enhanced by integrating domain-specific knowledge. This integration has give rise to informed machine learning, in contrast to studies that lack domain knowledge and treat all variables equally (uninformed machine learning). While the application of informed machine learning to bioinformatics and health informatics datasets has become more seamless, the likelihood of errors has also increased. To address this drawback, we present eight guidelines outlining best practices for employing informed machine learning methods in biomedical sciences. These quick tips offer recommendations on various aspects of informed machine learning analysis, aiming to assist researchers in generating more robust, explainable, and dependable results. Even if we originally crafted these eight simple suggestions for novices, we believe they are deemed relevant for expert computational researchers as well.

Introduction

Machine learning has become pervasive in a huge number of computational biology and medicine studies nowadays to address complicated problems being the backbone of the most novel research [1]. In fact, computational intelligence models help scientists understand complex biological processes [2], predict outcomes of a medical procedure [3,4], and support the design of new drugs [5]. Nevertheless, mistakes and bad practices in applying computational intelligence to biomedical data have become common, too [6–8].

The machine learning approaches can be nowadays categorized into two groups: informed and uninformed [9,10]. We call uninformed machine learning the models that do not make prior assumptions about the data set, meaning that they treat all the variables and the instances of the dataset in the same way, egalitarianly. These models do not take into account the biomedical knowledge that would beforehand highlight the role of a particular set of factors. On the contrary, we call informed machine learning the models which do take into account knowledge about the data set scientific subfield, during data collection and preparation (data

role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

pre-processing), during model development (in-processing), or during model correction and alignment (post-processing).

A study where a computational feature ranking phase is done on data of electronic medical records by treating all the variables in the same way [11], for example, can be considered an uninformed machine learning project. On the other hand, a bioinformatics study where three particular genes are known to be related to neuroblastoma and therefore treated with more importance compared to other genes [12] can be called informed machine learning.

It should be noted that the concepts of informed machine learning and uninformed machine learning have been previously introduced and developed in the field of artificial intelligence. These terms have their origins in background learning, which was first introduced in the 1980s. It is noteworthy that Art Samuel [13] was among the first to combine learning and domain knowledge-defined structure of the model, which could be considered the earliest example of informed machine learning, in 1959. Currently, there is a wealth of literature in this field that also proposes a historic perspective [14,15].

Before diving into the tips, let us first clarify what informed computational intelligence models are, when to use them, and help you recognize that you are likely already using them in many situations.

In order to understand what are the biologically and medically informed (partial-knowledge) computational intelligence models [16–18] and how to use them we have to start from the two most established approaches: the knowledge-based [10,16,19] (full-knowledge) and the data-driven based [11,20] (zero-knowledge) models.

Full-knowledge models strongly rely on humans and their comprehensive domain knowledge, employing a relatively small subset of available data and simple statistics primarily for validation rather than for model construction [10,19]. They often do not fully exploit all the possibly available data since some of them may be hard to exploit just based on domain knowledge [21]. Full-knowledge models are characterized by their predictability (as they are explainable by design) and adherence to physical plausibility, making them particularly suitable for biological and medical applications where the underlying mechanisms are well understood [22,23]. However, the effectiveness of these models is limited by the human capacity to conceptualize and manage complex biological systems, rendering them less flexible for novel or poorly understood phenomena [24–26].

Zero-knowledge models utilize large datasets to build and validate models without prior domain knowledge [27–30]. These models harness the computational power of modern technological infrastructures to analyze data, to identify patterns, and to make predictions that may not be immediately apparent to human researchers [31–34]. While these models excel in handling vast amounts of data and making generalized predictions, their outputs may not always be aligned with physical or biological plausibility or understandable and explainable, especially when extrapolating beyond the scope of the data; this limitation underscores the necessity for cautious interpretation of results, particularly in biological and medical settings where point wise accuracy and understanding the underlying mechanisms are crucial [35–39].

Partial-knowledge models represent a synthesis between the full- and zero-knowledge models capitalizing on the predictability, explainability, and plausibility of knowledge-based models while leveraging the data-processing capabilities of data-driven models. By integrating domain knowledge at varying levels, from data collection and feature engineering to the inclusion of approximate system models and model post-processing, partial-knowledge models strive for accuracy in both general and specific instances, enhancing their utility in biological and medical research for both interpolation and extrapolation tasks [10,16–19].

The integration of these modeling approaches within the biological and medical sciences can potentially enhance our understanding and treatment of complex diseases [40], facilitate

the discovery of new therapeutics [41], and contribute to personalized medicine [42]. By properly selecting and combining these methodologies, researchers can navigate the intricacies of biological systems and medical conditions, advancing the frontier of healthcare innovation [10,16–19].

Note that, some of the readers might have never heard about the terms informed machine learning and uninformed machine learning, and might read this manuscript thinking they are discovering something new. However, it is actually likely that they have already completed computational projects in both these areas in the past already.

The biomedical literature contains plenty of articles involving data-driven feature synthesis and selection, which could be categorized on informed machine learning (for example, [12,14]).

Both informed and uninformed machine learning have advantages and drawbacks and are error-prone. The Quick Tips series published articles on several computational aspects in the past, but none on this topic. We fill this gap by providing our eight quick tips for avoiding common mistakes and pitfalls when using informed machine learning in the biomedical sciences (Fig 1). We originally designed our recommendations for novices, but we believe they should be kept in mind by experts, too.

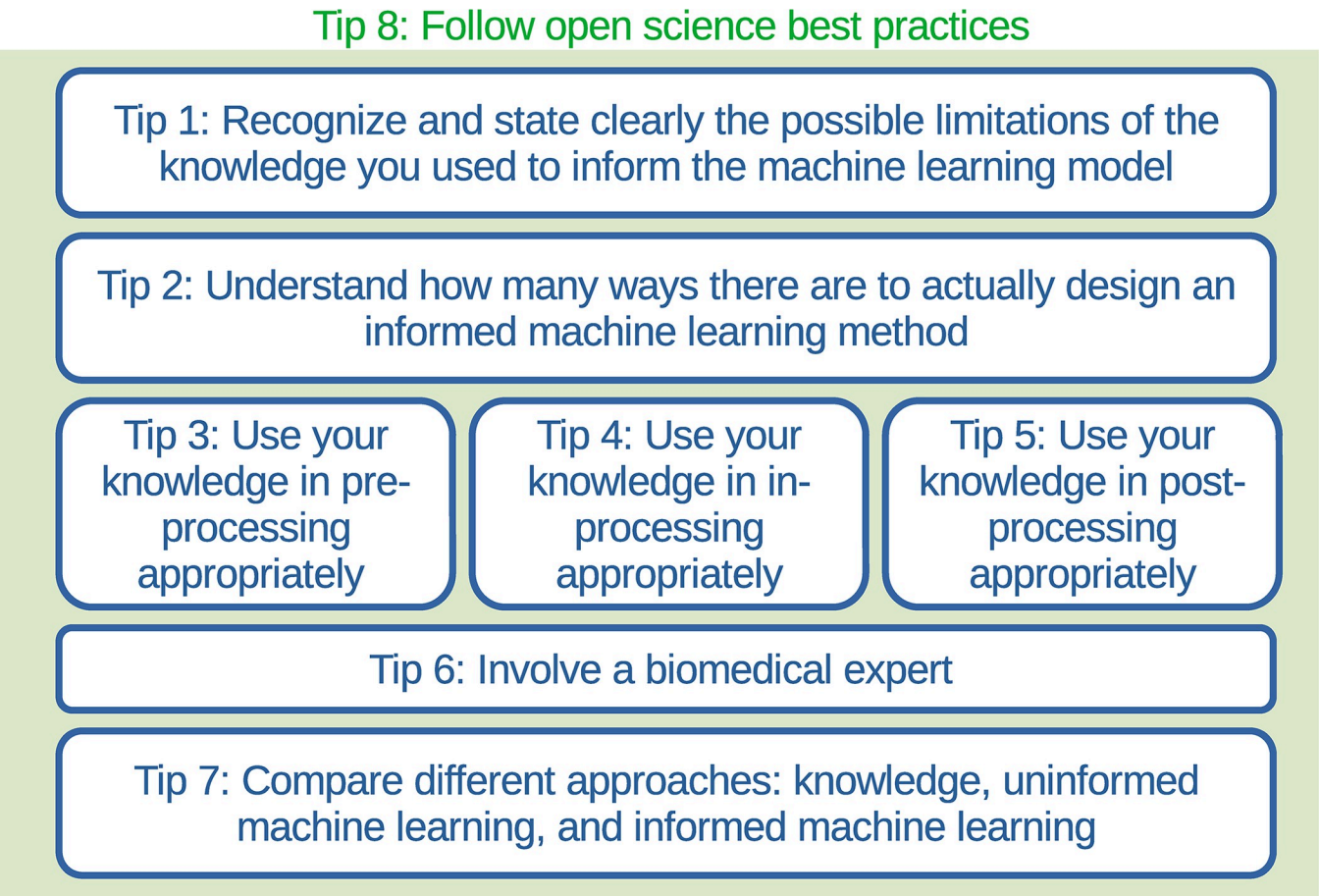

Fig 1. A schematic representation of the flowchart of the execution of our eight guidelines. The best practices for open science should be followed from the beginning to the end, and therefore, we represented them as the background of the whole process.

https://doi.org/10.1371/journal.pcbi.1012711.g001

Tip 1: Recognize and state clearly the possible limitations of the knowledge you used to inform the machine learning model

The usage of informed machine learning, as explained earlier, can bring several advantages to a scientific study. But it can, however, generate some problems as well.

Therefore, it is always important to keep in mind the limitations of this approach [43].

The informed machine learning approach, in fact, does not allow an agnostic inclusion of all the biomedical variables into a statistical model, and thus this selection sometimes can result being misleading [15]. For example, if the knowledge introduced by the informed model was wrong, outdated, obsolete, or misleading, it would corrupt the statistical model and generate inaccurate or inflated outcomes and results eventually. Sometimes, the information might be available as a general knowledge, but no specific feature related to that knowledge might have been annotated in the data set.

Occasionally, the biomedical experts and the computer scientists, although willing to collaborate, might not understand each other because of different experiences, knowledge, and jargon [44].

Other setbacks might happen because of the complexity of the model: including knowledge might seem useful, but if the statistical model became too complicated to be handled correctly, of course there would be no final benefit for the study.

Moreover, an informed machine learning model might fail to assimilate the knowledge introduced, and therefore might still work agnostically, even if the person who prepared the model thought they were preparing an informed machine learning algorithm [15,43].

Finally, past research has shown that adding additional knowledge to model inference might either increase or decrease the accuracy of the resulting models [45,46].

Tip 2: Understand how many ways there are to actually design an informed machine learning method

As described earlier, biologically and medically informed machine learning is an innovative approach that integrates domain-specific knowledge into data-driven models enhancing their performance, explainability, and plausibility [10,16,19]. This integration can be implemented at various stages of the machine learning pipeline, primarily categorized into pre-, in-, and post-processing methods.

Pre-processing is a critical step that involves preparing and transforming the data before it is fed into a machine learning model [47,48]. This stage is pivotal because it directly addresses the quality of input data, ensuring that the machine learning model has the best possible starting point. Techniques such as data cleaning [49], feature engineering [50,51], and data augmentation [52] fall under this category, where domain knowledge is leveraged to enhance the data set's relevance and quality. Furthermore, full-knowledge models that utilize domain knowledge to provide a first hint to be fed to or to be corrected by a machine learning model (that means, commonly referred as serial or parallel informed models) exemplify how domain knowledge can guide the machine learning model towards more accurate and relevant predictions [10]. Pre-processing can also improve explainability as machine learning model trained on a richer set of features and higher quality data can lead to simpler and then explainable models [17,53,54].

Pre-processing essentially capitalizes on domain expertise to navigate the machine learning model through the complex data landscape, minimizing the distance it needs to cover to generate valuable insights [55]. The pre-processing approach has been employed in several biomedical informatics studies in the past, especially when specific biomarkers were deemed more important than others before a computational phase [12].

In-processing involves the direct incorporation of domain knowledge into the computational intelligence model's learning process [56]. This method requires a deep integration of mathematical representations of domain insights (such as laws, trends, or constraints) into the learning algorithm itself [56] and demands a collaborative effort between domain experts and data scientists to modify the learning algorithm's structure [57]. This necessity could mean altering the functional form of the model [28,30] (for example, a particular architecture of a neural network), introducing specific constraints [58], or embedding regularizers [59] to maintain the model's desirable properties like convexity [60] and differentiability [61]. The objective is to steer the model's learning mechanism in a way that it not only benefits from domain knowledge but also enhances its predictive accuracy on a granular level, beyond average performance metrics [62]. By selection the proper model in-processing can also deal with the trade-off between accuracy and explainability [53]. The in-processing approach has been employed in multiple biomedical informatics studies in the past [10,56].

Post-processing focuses on refining the machine learning model's outputs to ensure they align with domain knowledge and expectations [63,64]. This stage does not modify the machine learning model itself but adjusts its outputs through additional rules or models to enforce domain consistency [65]. This alignment can improve also explainability by, for example, forcing the model to be never too far from a full-knowledge model [17,53]. Techniques include using machine learning predictions as inputs to physical models for more controlled outcomes or applying logical rules to rectify inconsistencies in predictions [66,67] (for example, a prediction indicating cancer should not concurrently suggest healthiness). Post-processing is about leveraging the existing machine learning capabilities as-is and employing domain knowledge to contextualize and correct the model's predictions [68]. This approach aims to mitigate potential errors and align the model's outputs with domain-specific truths, requiring substantial domain understanding to implement effectively [69,70]. The biomedical informatics literature has plenty of studies reporting post-processing informed machine learning approaches [10,19].

Tip 3: Use your knowledge in pre-processing appropriately

Pre-processing plays a crucial role in enhancing the predictive model performance by leveraging domain knowledge to inform, a priori, the machine learning model effectively [9,71].

Pre-processing acts, a priori, in different ways:

i. Modifying the input (that is, the features) fed to the machine learning models;

ii. Modifying the data (that is, the observation or samples) used to train the machine learning models; and

iii. Guiding the selection of the type of machine learning algorithms.

Modifying the input means cleaning, cleansing, engineering, selecting, and reducing the inputs to remove inconsistencies and errors, and enrich the input to ensure that the information fed into the machine learning models is of high quality [49,72].

Modifying the input also means constructing serial or parallel biologically and medically informed machine learning. In serial and parallel informed machine learning, a potentially partial (meaning that is unable to take into account all the available data) full-knowledge model is available to make prediction [73].

Modifying the data set involves selecting and enriching the available data [74], not only by choosing the most appropriate data but also by designing experiments to collect this data if necessary [75], ensuring it accurately represents the phenomena under study [76]. Techniques for data fusion and data integration are a key in this context [77].

Experimental design for data collection is probably the most important phase of a successful machine learning-based research project or product as it allows to prevent to introduce spurious correlations [78] and to try to match the main hypothesis behind any machine learning algorithm, namely the data well represent the population [79].

Guiding the selection of the type of machine learning algorithm means guiding the selection of the algorithm functional form [30] (for instance, deep or shallow and for the deep the type of architecture), including transfer learning [80], the level of explainability of the model (from rule based model to deep models passing from linear models) [53], and the hyperparameters characterizing the machine learning algorithm [81,82]. This aspect is a pivotal part of the process and should be informed by the specific characteristics of the pre-processed and enriched data set and the physiological mechanisms/principles and by the domain knowledge underpinning the problem space [83].

For example, if we are in a safety-critical situation, it is better to use a fully interpretable model [53], even if less accurate. If we deal with images, convolutions are the best choice, while transformers are the way to go if we have to deal with natural language [28]. If we have a lot of structured data, deep models are probably the best choice, while for medium or small cardinality unstructured data sets, shallow models are the optimal choice [30]. This mix of domain knowledge and experience can make a difference in delivering an effective biomedically informed machine learning study.

Tip 4: Use your knowledge in in-processing appropriately

In-processing plays a crucial role in enhancing the predictive model performance by leveraging domain knowledge of a potentially partial, but well mathematically encoded, full-knowledge model to inform the learning process of a zero-knowledge effectively [56].

In-processing involves the integration of domain-specific laws and principles directly into the model training process, especially to ensure that the models adhere closely to known scientific knowledge [84]. When certain biological laws or medical principles are known, these can be used to guide the model in several ways [69]. One approach is to ensure the model's predictions do not deviate significantly from these laws by incorporating a regularization term that penalizes deviations from the expected physical behavior [68]. This regularization can help in maintaining the model's fidelity to the biological laws or medical principles, such as the relationship between specific inputs and outputs, ensuring that if an input value increases, the output adjusts in a biological or medical consistent manner [68,69]. Moreover, when dealing with complex systems where simulators exist but are too slow for practical use, surrogate models can be developed [84]. These models aim to mimic the simulator's outputs while being computationally efficient, thus requiring the model to accurately capture the underlying physical relationships [56]. Incorporating biological laws or medical principles into machine learning does not always require exhaustive or precise details; even hints or partial knowledge about the physical system can be beneficial [10,56]. Other methods like the ones based on reasoning like inductive logic programming [85], neuro-symbolic approaches [67,86], and learning constrained models [58] have also shown to achieve good practical results.

In summary, the in-processing strategy within biologically and medically informed machine learning represents a paradigm shift toward developing machine learning models that are informed by and compliant with the biological laws or medical principles. This approach significantly contributes to creating models that are not only predictive but also interpretative and aligned with the real-world phenomena they aim to simulate, thereby bridging the gap between data-driven insights and biological laws or medical principles.

Tip 5: Use your knowledge in post-processing appropriately

Post-processing leverages domain knowledge (for example, a potentially, partial but still well-mathematically encoded, full-knowledge model) to align the output of a zero-knowledge model [63].

Post-processing in physics-informed models focuses on refining predictions of a zero-knowledge model to ensure they align with domain-specific knowledge, constraints, and practical considerations [64]. This step is crucial for enforcing certain characteristics and relationships that must be present in the predictions. For instance, in an autonomous system to diagnose different types of cancers, if a model predicts the type of cancer, it must recognize hard constraints such as the hierarchies present in the diseases, namely, if we predict cancer in the lung we also have to predict lung problems [66,67].

Furthermore, post-processing involves adjusting predictions to ensure they do not diverge excessively from established practices, like the dose of chemotherapy, thereby ensuring that the model's recommendations are practical and implementable within current protocols. This step not only enhances the model's reliability but also its acceptance among practitioners [63]. Ensuring that the model's decisions are in harmony with domain knowledge can be straightforward if the model is inherently explainable [53]. For models that are not easily interpretable, techniques such as feature importance analysis can provide global explainability, offering insights into what the model considers important across all decisions [53]. For local explainability—understanding individual predictions—tools can be employed to break down the decision-making process for specific cases [53].

Eventually, one can decide to build full-knowledge model, then interpretable and controllable, that require some inputs that are not easy to measure or estimate: in this case zero-knowledge model might be of support and more easily controllable, since their estimate can be verified and controlled by the subsequent full-knowledge model toward final biological and medical informed machine learning model [64].

Tip 6: Involve a biomedical expert

As explained in other Quick Tips articles [87], the success of an interdisciplinary scientific project relies massively on the involvement of domain experts for all the scientific fields involved. So, if your study is on bioinformatics, we suggest you to involve a wet-lab biologist, and if your study is on medical informatics, we advise you contact a medical doctor. These biomedical experts need to be contacted and included in two main phases: at the beginning of your project (when you are defining the scientific question) and at the end (when the results are ready and their implications need to be discerned) [88].

Therefore, if you work at a university, we recommend that you contact a biologist in the biology department or a medical doctor in the medicine department [44]. If there are no biomedical experts in your organization, we recommend that you look for someone online, on forums such as Reddit, StackExchange, or similar.

The information given by the biomedical expert will be pivotal: they will provide insights for the scientific question definition and for the result understanding that will be invaluable and that will enrich your study.

Of course, it would be even better if you could have the support of a biomedical researcher throughout the whole project and not just at the start and at the end. This hope, however, can be too optimistic knowing the busy schedule of medical doctors and wet-lab biologists [44].

Tip 7: Evaluate and compare different approaches: Knowledge, uninformed machine learning, and informed machine learning

One of the best practices of computational projects is to use different methods to see if similar results are found. The same approach can be employed in a biomedical study: therefore, when you have a biomedical data set ready to be analyzed, we suggest you to process it through a knowledge approach (using common knowledge about the scientific subfield and standard statistics about the data set, without machine learning [19]), a uninformed machine learning approach (also called data-driven or zero-knowledge [11,20]), and an informed machine learning approach (also called partial-knowledge [19]).

A knowledge strategy that includes only information about the scientific domain and involves no computational intelligence can be carried out through traditional statistics [89], by checking against common knowledge: for example, if a specific clinical factor is known to be prognostic for the disease of the data set, one can double-check if the results obtained through the informed or uninformed machine learning approaches confirm it or not.

For instance, in a previous study on a data set of electronic health records with chronic kidney disease, we utilized a uninformed machine learning approached which detected age, estimated glomerular filtration rate (eGFR), and creatinine as most diagnostic factors [90]. Afterwards, we double-checked our discoveries in the scientific literature (knowledge approach) to see if they could be confirmed or not. That is what we suggest you to do.

The results of informed machine learning studies [12] should be checked against common knowledge on the disease investigated, too. Comparison of results can lead to interesting insights about the data (for example, if more data might be needed to complete the study), or about the methods (for example, if the results contradict common knowledge), or about new discoveries.

Regarding evaluation, past research has shown that adding additional knowledge to model inference might either increase or decrease the accuracy of the resulting models, but it consistently aids in generating explainable models [45,46].

Tip 8: Follow open science best practices

Even for informed machine learning, we promote the usage of best practices for open science: open source software code, open data release, and open access publication [91].

If you have the chance to decide which programming language to use for your informed machine learning project, we strongly suggest to pick an open source one, such as Python or R. This way, you will be able to share your software code with any collaborator at any time and, if you publish your software code openly later on GitHub or GitLab, anyone will be able to use it. These practices would ensure the possibility to reproduce and replicate your computational results, and would allow other researchers around the world to start new, similar scientific projects if they want.

Regarding data, we recommend that you share your data online in public, open repositories, if you are authorized to do so. There are several open repositories for bioinformatics data and medical data where you can release your data set (Gene Expression Omnibus (GEO), ArrayExpress, Sequence Read Archive (SRA), and the Cancer Genome Atlas (TCGA), the Cancer Imaging Archive (TCIA), and PhysioNet, for example) and for any data type (Figshare, Zenodo, and University of California Irvine Machine Learning Repository).

We advise you to openly publish both the raw data and the preprocessed data you used for your analysis. Of course, the privacy of data's patients need to be preserved: make sure that all the data are anonymous, deidentified, and unidentifiable.

Moreover, as explained in other Quick Tips articles [88,92,93], we suggest to use these online resources to look for an alternative data set of the same type and of the same disease of the primary cohort dataset that you analyzed in your study. If you found one, repeat your analysis on it and see if your scientific discoveries are confirmed there.

Finally, if you have a say on which scientific journal to choose for your paper submission, we suggest to pick an open access one: publishing an open access article, in fact, would make it readable and available to anyone in the world, and also let your study have a bigger impact on the scientific community.

Conclusions

Informed machine learning has become popular in several biomedical studies nowadays, thanks to the large availability of computational resources and the spread of knowledge about computational intelligence. Even if it has become easier to apply informed machine learning, it has become easier to make mistakes, too: bad practices and pitfalls, if not carefully handled, that can produce negative consequences on the final results of the study. In this manuscript, we propose eight simple guidelines for avoiding common mistakes and inaccuracies in studies involving informed machine learning phases.

We believe our eight recommendations can help researchers produce more stable and reliable results in any biomedical study.

Author Contributions

Conceptualization: Luca Oneto, Davide Chicco.

Formal analysis: Luca Oneto.

Investigation: Luca Oneto, Davide Chicco.

Methodology: Luca Oneto, Davide Chicco.

Project administration: Davide Chicco.

Resources: Davide Chicco.

Supervision: Luca Oneto.

Writing – original draft: Luca Oneto, Davide Chicco.

Writing – review & editing: Luca Oneto, Davide Chicco.

References

1. Winslow RL, Trayanova N, Geman D, Miller MI. Computational medicine: translating models to clinical care. Sci Transl Med. 2012; 4(158):158rv11–158rv11. https://doi.org/10.1126/scitranslmed.3003528 PMID: 23115356
2. Karimzadeh M, Cavazos TB, Chen NC, Tbeileh NK, Siegel D, Momen-Roknabadi A, et al. Beyond detection: AI-based classification of breast cancer invasiveness using cell-free orphan non-coding RNAs. Cancer Res. 2024; 84(6 Supplement):3678–3678.
3. Haleem A, Javaid M, Khan IH. Current status and applications of artificial intelligence (AI) in medical field: an overview. Curr Med Res Pract. 2019; 9(6):231–237.
4. Weller GB, Lovely J, Larson DW, Earnshaw BA, Huebner M. Leveraging electronic health records for predictive modeling of post-surgical complications. Stat Methods Med Res. 2018; 27(11):3271–3285. https://doi.org/10.1177/0962280217696115 PMID: 29298612
5. Ceddia G, Pinoli P, Ceri S, Masseroli M. Matrix factorization-based technique for drug repurposing predictions. IEEE J Biomed Health Inform. 2020; 24(11):3162–3172. https://doi.org/10.1109/JBHI.2020.2991763 PMID: 32365039

6. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform. 2021; 153:104510. https://doi.org/10.1016/j.ijmedinf.2021.104510 PMID: 34108105
7. Makin TR, Orban de Xivry JJ. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. Elife. 2019; 8:e48175. https://doi.org/10.7554/eLife.48175 PMID: 31596231
8. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012; 55(10):78–87.
9. von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng. 2021; 35(1):614–633.
10. Leiser F, Rank S, Schmidt-Kraepelin M, Thiebes S, Sunyaev A. Medical informed machine learning: a scoping review and future research directions. Artif Intell Med. 2023; 145:102676. https://doi.org/10.1016/j.artmed.2023.102676 PMID: 37925206
11. Chicco D, Haupt R, Garaventa A, Uva P, Luksch R, Cangelosi D. Computational intelligence analysis of high-risk neuroblastoma patient health records reveals time to maximum response as one of the most relevant factors for outcome prediction. Eur J Cancer. 2023; 193:113291. https://doi.org/10.1016/j.ejca.2023.113291 PMID: 37708628
12. Chicco D, Sanavia T, Jurman G. Signature literature review reveals AHCY, DPYSL3, and NME1 as the most recurrent prognostic genes for neuroblastoma. BioData Mining. 2023; 16(1):7. https://doi.org/10.1186/s13040-023-00325-1 PMID: 36870971
13. Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev. 1959; 3 (3):210–229.
14. Mao L, Wang H, Hu LS, Tran NL, Canoll PD, Swanson KR, et al. Knowledge-informed machine learning for cancer diagnosis and prognosis: a review. arXiv preprint. 2024;arXiv:2401.06406.
15. Hao Z, Liu S, Zhang Y, Ying C, Feng Y, Su H, et al. Physics-informed machine learning: a survey on problems, methods and applications. arXiv preprint. 2022;arXiv:2211.08064.
16. Oberste L, Rüffer F, Aydingül O, Rink J, Heinzl A. Designing user-centric explanations for medical imaging with informed machine learning. In: International Conference on Design Science Research in Information Systems and Technology; 2023. p. 470–484.
17. Oberste L, Heinzl A. User-centric explainability in healthcare: a knowledge-level perspective of informed machine learning. IEEE Trans Artif Intell. 2023; 4(4):840–857.
18. Khayal IS, O'Malley AJ, Barnato AE. Clinically informed machine learning elucidates the shape of hospice racial disparities within hospitals. NPJ Digit Med. 2023; 6(1):190. https://doi.org/10.1038/s41746-023-00925-5 PMID: 37828119
19. Johnson M, Albizri A, Harfouche A, Fosso-Wamba S. Integrating human knowledge into artificial intelligence for complex and ill-structured problems: informed artificial intelligence. Int J Inf Manag. 2022; 64:102479.
20. Chiu YL, Jhou MJ, Lee TS, Lu CJ, Chen MS. Health data-driven machine learning algorithms applied to risk indicators assessment for chronic kidney disease. Risk Management and Healthcare Foreign Policy. 2021; p. 4401–4412. https://doi.org/10.2147/RMHP.S319405 PMID: 34737657
21. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. In: AMIA Annual Symposium Proceedings. vol. 2013. American Medical Informatics Association; 2013. p. 1472.
22. Bernasconi A, Zanga A, Lucas PJ, Stella MSF. Towards a transportable causal network model based on observational healthcare data. arXiv preprint. 2023;arXiv:2311.08427.
23. Zanga A, Bernasconi A, Lucas PJ, Pijnenborg H, Reijnen C, Scutari M et al. Risk assessment of lymph node metastases in endometrial cancer patients: a causal approach. arXiv preprint. 2023; arXiv:2305.10041.
24. Grosch E. Reply to "Ten simple rules for getting published". PLoS Comput Biol. 2007; 3(9):e190. https://doi.org/10.1371/journal.pcbi.0030190 PMID: 17907799
25. Altman DG. Poor-quality medical research: what can journals do? JAMA. 2002; 287(21):2765–2767. https://doi.org/10.1001/jama.287.21.2765 PMID: 12038906
26. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005; 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
27. Foster D. Generative Deep Learning. Sebastopol, California, USA: O'Reilly Media; 2022.
28. Aggarwal CC. Neural Networks and Deep Learning: A Textbook. Cham, Switzerland: Springer; 2023.
29. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambrdige, Massachusetts, USA: MIT press; 2016.

30. Shalev-Shwartz S, Ben-David S. Understanding machine learning: from theory to algorithms. Cambridge, England, United Kingdom: Cambridge University Press; 2014.
31. Hinkson IV, Madej B, Stahlberg EA. Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. Front Pharmacol. 2020; 11:770. https://doi.org/10.3389/fphar.2020.00770 PMID: 32694991
32. Butler D. Tomorrow's world: technological change is accelerating today at an unprecedented speed and could create a world we can barely begin to imagine. Nature. 2016; 530(7591):398–402.
33. Melo MCR, Maasch JRMA, De La Fuente-Nunez C. Accelerating antibiotic discovery through artificial intelligence. Commun Biol. 2021; 4(1):1050. https://doi.org/10.1038/s42003-021-02586-0 PMID: 34504303
34. Monroe D. Accelerating AI. Commun ACM. 2022; 65(3):15–16.
35. Di Nucci E. Should we be afraid of medical AI? J Med Ethics. 2019; 45(8):556–558. https://doi.org/10.1136/medethics-2018-105281 PMID: 31227547
36. Chin-Yee B, Upshur R. Three problems with big data and artificial intelligence in medicine. Perspect Biol Med. 2019; 62(2):237–256. https://doi.org/10.1353/pbm.2019.0012 PMID: 31281120
37. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022; 28(1):31–38. https://doi.org/10.1038/s41591-021-01614-0 PMID: 35058619
38. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intell. 2020; 2(11):665–673.
39. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023; 24(2):125–137. https://doi.org/10.1038/s41576-022-00532-2 PMID: 36192604
40. Martínez-García M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine. Front Med. 2022; 8:784455. https://doi.org/10.3389/fmed.2021.784455 PMID: 35145977
41. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. J Chem. 2013; 5:1–9. https://doi.org/10.1186/1758-2946-5-30 PMID: 23800010
42. Siddiq M. Integration of machine learning in clinical decision support systems. Eduvest-Journal of Universal Studies. 2021; 1(12):1579–1591.
43. Fuks O, Tchelepi HA. Limitations of physics informed machine learning for nonlinear two-phase transport in porous media. J Mach Learn Model Comput. 2020; 1(1).
44. Chicco D, Jurman G. Ten simple rules for providing bioinformatics support within a hospital. BioData Mining. 2023; 16(1):6. https://doi.org/10.1186/s13040-023-00326-0 PMID: 36823520
45. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. Electronics. 2019; 8(8):832. https://doi.org/10.3390/electronics8080832
46. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A. 2019; 116(44):22071–22080. https://doi.org/10.1073/pnas.1900654116 PMID: 31619572
47. Li Q, Liu C, Oster J, Clifford GD. Signal processing and feature selection preprocessing for classification in noisy healthcare data. Machine Learning for Healthcare Technologies. 2016; 2(33):2016.
48. Tam S, Tsao MS, McPherson JD. Optimization of miRNA-seq data preprocessing. Brief Bioinform. 2015; 16(6):950–963. https://doi.org/10.1093/bib/bbv019 PMID: 25888698
49. Ilyas IF, Chu X. Data Cleaning. Redwood City, California, USA: Morgan & Claypool; 2019.
50. Chicco D, Oneto L, Tavazzi E. Eleven quick tips for data cleaning and feature engineering. PLoS Comput Biol. 2022; 18(12):e1010718. https://doi.org/10.1371/journal.pcbi.1010718 PMID: 36520712
51. Duboue P. The art of feature engineering: essentials for machine learning. Cambridge, England, United Kingdom: Cambridge University Press; 2020.
52. Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. Array. 2022; 16:100258.
53. Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res. 2021; 70:245–317.
54. Yang W, Wei Y, Wei H, Chen Y, Huang G, Li X, et al. Survey on explainable AI: From approaches, limitations and applications aspects. Human-Centric Intelligent Systems. 2023; 3(3):161–188.
55. Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. J Eng Appl Sci. 2017; 12(16):4102–4107.
56. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021; 3(6):422–440.

57. Kumar P, Sharma M. Data, machine learning, and human domain experts: none is better than their collaboration. Int J Hum Comput Interact. 2022; 38(14):1307–1320.

58. Gori M, Betti A, Melacci S. Machine learning: a constraint-based approach. Amsterdam, the Netherlands: Elsevier; 2023.

59. Oneto L, Navarin N, Biggio B, Errica F, Micheli A, Scarselli F, et al. Towards learning trustworthily, automatically, and with guarantees on graphs: an overview. Neurocomputing. 2022; 493:217–243.

60. Bartlett PL, Jordan MI, McAuliffe JD. Convexity, classification, and risk bounds. J Am Stat Assoc. 2006; 101(473):138–156.

61. Hernández A, Millerioux G, Amigó JM. Differentiable programming: generalization, characterization and limitations of deep learning. arXiv preprint. 2022;arXiv:2205.06898.

62. Jayatilake SMDAC, Ganegoda GU. Involvement of machine learning tools in healthcare decision making. J Healthc Eng. 2021;2021. https://doi.org/10.1155/2021/6679512 PMID: 33575021

63. Halder S, Yamasaki J, Acharya S, Kou W, Elisha G, Carlson DA, et al. Virtual disease landscape using mechanics-informed machine learning: application to esophageal disorders. Artif Intell Med. 2022; 134:102435. https://doi.org/10.1016/j.artmed.2022.102435 PMID: 36462900

64. Magni M, Sutanudjaja EH, Shen Y, Karssenberg D. Global streamflow modelling using process-informed machine learning. J Hydroinformatics. 2023; 25(5):1648–1666.

65. Sanchez-Garcia R, Gomez-Blanco J, Cuervo A, Carazo JM, Sorzano COS, Vargas J. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. Commun Biol. 2021; 4(1):874. https://doi.org/10.1038/s42003-021-02399-1 PMID: 34267316

66. Giunchiglia E, Stoian MC, Lukasiewicz T. Deep learning with logical constraints. In: International Joint Conference on Artificial Intelligence; 2022. p. 5478–5485.

67. Giunchiglia E, Imrie F, van der Schaar M, Lukasiewicz T. Machine learning with requirements: a manifesto. arXiv preprint. 2023;arXiv:2304.03674.

68. Huang J, Yan H, Li J, Stewart HM, Setzer F. Combining anatomical constraints and deep learning for 3-D CBCT dental image multi-label segmentation. In: Proceedings of ICDE 2021 –the 37th IEEE International Conference on Data Engineering. IEEE; 2021. p. 1–6.

69. Huynh PK, Setty A, Phan H, Le TQ. Probabilistic domain-knowledge modeling of disorder pathogenesis for dynamics forecasting of acute onset. Artif Intell Med. 2021; 115:102056. https://doi.org/10.1016/j.artmed.2021.102056 PMID: 34001316

70. Azmat M, Tu E, Branch KR, Alessio A. Machine learned versus analytical models for estimation of Fractional Flow Reserve (FFR) from CT-derived information. In: Gimi BS, Krol A, editors. Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging; 2021. p. 212–217.

71. Cheng CY, Li Y, Varala K, Bubert J, Huang J, Kim GJ, et al. Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. Nature IDAA Commun. 2021;12(1). https://doi.org/10.1038/s41467-021-25893-w PMID: 34561450

72. He B, Zhu R, Yang H, Lu Q, Wang W, Song L, et al. Assessing the impact of data preprocessing on analyzing next generation sequencing data. Front Bioeng Biotechnol. 2020; 8:817. https://doi.org/10.3389/fbioe.2020.00817 PMID: 32850708

73. Kroll A. Grey-box models: concepts and application. New Frontiers in Computational Intelligence and its Applications. 2000; 57:42–51.

74. Aupetit M. Nearly homogeneous multi-partitioning with a deterministic generator. Neurocomputing. 2009; 72(7–9):1379–1389.

75. Newman A, Bavik YL, Mount M, Shao B. Data collection via online platforms: Challenges and recommendations for future research. Appl Psychol. 2021; 70(3):1380–1402.

76. Kilkenny MF, Robinson KM. Data quality:"Garbage in–garbage out". Health Inf Manag J. 2018; 47 (3):103–105.

77. Zhang Y, Sheng M, Liu X, Wang R, Lin W, Ren P, et al. A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. Health Inf Sci Syst. 2022; 10(1):22. https://doi.org/10.1007/s13755-022-00183-x PMID: 36039096

78. Haig BD. What is a spurious correlation? Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences. 2003; 2(2):125–132.

79. Chen RJ, Chen TY, Lipkova J, Wang JJ, Williamson DF, Lu MY, et al. Algorithm fairness in AI for medicine and healthcare. arXiv preprint. 2021;arXiv:2110.00603.

80. Kora P, Ooi CP, Faust O, Raghavendra U, Gudigar A, Chan WY, et al. Transfer learning techniques for medical image analysis: A review. Biocybernetics and Biomedical Engineering. 2022; 42(1):79–107.

81. Lindauer M, Hutter F. Best practices for scientific research on neural architecture search. J Mach Learn Res. 2020; 21(243):1–18.

82. Oneto L. Model selection and error estimation in a nutshell. Cham, Switzerland: Springer; 2020.
83. Guidoboni G, Zou D, Lin M, Nunez R, Rai R, Keller J, et al. Physiology-informed machine learning to enable precision medical approaches of intraocular pressure and blood pressure management in glaucoma. Invest Ophthalmol Vis Sci. 2022; 63(7):2293–2293.
84. Azmat M, Tu E, Branch KR, Alessio AM. Machine learned versus analytical models for estimation of fractional flow reserve from CT-derived information. In: Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging. vol. 11600. SPIE; 2021. p. 212–217.
85. Siromoney A, Raghuram L, Siromoney A, Korah I, Prasad GNS. Inductive logic programming for knowledge discovery from MRI data. IEEE Eng Med Biol Mag. 2000; 19(4):72–77. https://doi.org/10.1109/51.853484 PMID: 10916735
86. Kang T, Turfah A, Kim J, Perotte A, Weng C. A neuro-symbolic method for understanding free-text medical evidence. J Am Med Inform Assoc. 2021; 28(8):1703–1711. https://doi.org/10.1093/jamia/ocab077 PMID: 33956981
87. Chicco D, Agapito G. Nine quick tips for pathway enrichment analysis. PLoS Comput Biol. 2022; 18(8): e1010348. https://doi.org/10.1371/journal.pcbi.1010348 PMID: 35951505
88. Cisotto G, Chicco D. Ten quick tips for clinical electroencephalographic (EEG) data acquisition and signal processing. PeerJ Comput Sci. 2024; 10:e2256. https://doi.org/10.7717/peerj-cs.2256 PMID: 39314688
89. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. Hoboken, New Jersey, USA: Wiley; 2018.
90. Chicco D, Lovejoy CA, Oneto L. A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease. IEEE Access. 2021; 9:165132–165144.
91. Markowetz F. Five selfish reasons to work reproducibly. Genome Biol. 2015; 16:1–4.
92. Chicco D, Karaiskou AI, De Vos M. Ten quick tips for electrocardiogram (ECG) signal processing. PeerJ Comput Sci. 2024; 10:e2295. https://doi.org/10.7717/peerj-cs.2295 PMID: 39314696
93. Bonnici V, Chicco D. Seven quick tips for gene-focused computational pangenomic analysis. BioData Mining. 2024; 17(1):28. https://doi.org/10.1186/s13040-024-00380-2 PMID: 39227987