Editorial

Int Neurourol J 2024;28(1):1-3
https://doi.org/10.5213/inj.2424edi02
pISSN 2093-4777 · eISSN 2093-6931

INTERNATIONAL NEUROUROLOGY JOURNAL
INJ
OPEN ACCESS

# Coping With Methodology for Research in the Age of Artificial Intelligence: A Reader's Guide to "Prediction of Prostate Cancer Risk Stratification Based on a Nonlinear Transformation Stacking Learning Strategy"

Jin Wook Kim[1,2] https://orcid.org/0000-0003-4157-9365

[1]Department of Medical Informatics, Chung-Ang University, Seoul, Korea
[2]Department of Urology, Chung-Ang University Gwangmyeong Hospital, Gwangmyeong, Korea

Email: jinwook@cau.ac.kr

Despite difficulties for seasoned investigators to cope with the ever-increasing laundry list of methodologies to master, the advent of machine learning (ML) and in the general interest of the public for artificial intelligence (AI) has rocketed proportional expectations in its performance in fields outside computer science as well. However, like other recent yet more established modalities, investigation with ML/AI tools should be perceived not as a blanketed enigma in a black box, but rather as a tangible statistical model with strengths and weaknesses like any other, albeit different and novel.

As when research had to adopt for more stringent measures in engaging clinical research with larger populations, randomization, stratification and proportional enrolment, or as when research begun to absorb meta-analytical outcomes while applying rigorous criteria for inclusion, or even as basic research ever expanded its investigative tools to incorporate molecular, genetic, epigenetic, metabolomic, optogenetic methodologies, we must now familiarize ourselves to a professional level with the parlance of ML/AI.

This is not as daunting a task as it may seem, and we encourage both budding and experienced researchers to understand the nomenclature and methodologies of the new field of ML/AI incorporated scientific research. The perception of its discouraging incomprehensibility bordering on fantasy resides primarily in recent strides of more consumer-friendly examples, even though, in a sense of fundamental principle these are less complex and more data heavy than the research methodologies we are required to adapt to.

Without further preamble, it is sufficient to state that understanding of ML methods has been of significant interest, are now important and in the future necessary. As it is not something that is going away, it is best to start understanding it, now.

## The Problem of Overfitting

Before engaging in the details, it is important to understand ML methodologies as a form of statistical modelling. The outcomes are ultimately interpretable as predictability, viz a viz, sensitivity, specificity, false negatives/positives, and probabilities. While some studies may employ continuous outcomes within the process, ultimately as a research outcome it is easier to package the entire research strategy within as a multivariate generalized linear process with logistic regression outcomes.

Fig. 2 of Cao et al. [1] is the basic map that will guide the reader through understanding the methodology. Most ML/AI manuscripts provide similar roadmaps of research, albeit of varying understandability. We can first focus on the end result, where we can see the familiar features of the receiver operating characteristic (ROC) curve. As an aside, the ROC was created

in the midst of the second world war to understand the outcomes of the new scientific methodology at the time, radar; the scenario is eerily familiar.

To derive this outcome, the researchers have employed 3 comparative methods (a, b, c). Option a, termed "standard stacking" can be considered as the negative control method. However, unlike traditional methodologies in the clinical or laboratory settings where a method applied to a group is irrevocable, thus limiting the investigator to surmise the counterfactual outcomes at best, in silico methods (including, but not limited to, ML) is not constrained by irrevocability of effect and counterfactuals can be applied ad infinitum. This, of course, results in the fundamental weakness of ML research, overfitting.

The concept of overfitting is a core drawback when interpreting results from ML method-based research. In conventional research, a comparable but far less debilitating level of detraction would be overestimated P-values with overly cumbersome populations [2]. Overfitting is similar as it provide hyper resolution to trivial details exaggerating the effects while diminishing its reproducibility, i.e., through methods which enhance the outcomes the method ultimately binds itself to only that particular scenario. ML methodologies are capable of taking that to the extreme. This is not a mere apocryphal tale; most notably, AI aided diagnosis of diabetic retinopathy, which initially was touted as possibly replacing retina specialists, failed utterly with new patient sets or on hands applications [3]. As such, alarming ROC curves as seen in Fig. 5A of Cao et al. [1] is par for the course in ML. However, the realistically lower ROC curves in Fig. 5B can still worsen when applied in the ground truth.

## Stacking

Stacking is a form of ensemble learning [4]. Basically, ensemble learning is a meta methodology that combines several base training models in various ways to produce a congregated outcome. Despite its unfamiliar nomenclature, ensemble learning is not new to the hardened data scientist. The most basic method is bootstrapping, where a large sample is spliced to produce smaller samples at random (its elements are reusable and possibly chosen multiple times through each iteration) and trained. The results of multiple sample trainings are then aggregated; this is called bagging (shortened from bootstrap aggregating). Boosting is an altered form of bagging, where a subsample of the original data is trained, but each failed predictions of the iterations are then subsequently emphasized in retraining to further adapt the model. Stacking is another method of this line of investigation, where multiple different methods are used to train base models. The initial outcomes of these base models are then used to train a meta model.

## Nonlinear Transformations

Nonlinear transformations, as presented in the paper in Fig 2 and subsection "standard stacking with nonlinear transformation" are simply a series of transformation functions. These functions are most popularly used as activation functions for nodes composing neural networks for deep learning, primarily to minimize the processing burden. One of the most popular neural network function rectified linear unit (ReLU) has a straightforward nonlinear relation of max (0, $x$) (i.e., if input is smaller than 0, then 0; if larger than zero, input is output). Smoothened versions of ReLU are GeLU (Gaussian error linear unit) or softmax, which have been used in this paper. Of course, the most familiar one is the naturally occurring function of enzymes and chemical reactions, the sigmoid function. Other transformations are of similar nature, aimed to provide the overall meta model egalitarian selection without bias over the base models to best fit the overall target.

As such, these nonlinear transformations can provide outputs of the base models to plug into the meta model, which is shown to be a logistic regression, a model best suited to accept such inputs.

## Occam's Razor in the Age of AI

To summarize what we have taken into account, the paper by Cao et al. [1] is an exercise in stacking, albeit in a novel way of utilizing nonlinear transformations. Qualitatively, it could be seen as an attempt to utilize and balance out multiple high-yield statistical methodologies.

The study, ultimately, is a study that investigates 197 patients, which included 59 low-risk, 48 intermediate-risk, and 90 high-risk patients. By conventional standards, 197 patients are not a large number, especially not something to derive stratification of risk from. Yet, judging from the ROC curve, the investigation achieved not only statistical significance, but also overfitting. The overfitting of the training model, and the moderated outcome of the validation model shows that the study was sufficiently powered. As such one could see that ML could be used to derive scientific results on par with older methodologies.

Yet, despite the convoluted stacking and sampling, the study has been overfitted. And that overfitting has fallen far short of the first validation within paper. This is the fundamental cau-

tion towards ML in the use as a tool of investigation. We do not know exactly what weights were given, why and how, within the meta model to derive this conclusion. This is not a criticism against the researchers. It is a warning of caution against the ML itself, that the unknown components that adjusted the weights appropriately to overfit the training model could incorporate unseen or unimputed influences within the model. Despite not being within the variable, the nature of the population, the proclivities of the investigators, the weather, the political situation, the price of oil and other mitigating unseen factors can be detected in absentia by ML. These factors, called latent variables, are unseen variables that are not included in the data, yet may emerge as an influence of the outcome. And despite notable quandaries where they have confused major AI models, they have not yet surfaced in the mainstream research as a conscious and active prevention target as Bonferroni correction is to false positives. While this reasoning may seem convoluted, there is one rule of science that cuts through all this frivolous discussion of overfitting versus latent variables, and that is the Principle of Parsimony, otherwise known as Occam's Razor [5].

Specifically, how one should maintain Occam's Razor in the age of ML is beyond the scope of this editorial, but it is evident that it should.

• **Conflict of Interest:** No potential conflict of interest relevant to this article was reported.

## REFERENCES

1. Cao X, Fang Y, Yang C, Liu Z, Xu G, Jiang Y, et al. Prediction of prostate cancer risk stratification based on a nonlinear transformation stacking learning strategy. Int Neurourol J 2024;28:33-43.
2. Lin M, Lucas HC, Shmueli G. Research commentary: too big to fail: large samples and the p-value problem. Inform Syst Res 2013;24:906-17.
3. Senapati A, Tripathy HK, Sharma V, Gandomi AH. Artificial intelligence for diabetic retinopathy detection: a systematic review. Inform Med Unlocked 2024;45:101445.
4. Polikar R. Ensemble learning. In: Zhang C, Ma Y, editors. Ensemble machine learning: methods and applications. Berlin (Germany): Springer Nature; 2012. p. 1-34.
5. Information and likelihood theory: a basis for model selection and inference. In: Burnham KP, Anderson DR, editors. Model selection and multimodel inference. New York: Springer; 2002. p. 49-97.