Journal of Applied Statistics & Machine Learning
Volume 1; Number 2; 2022 : pp. 117-126

CHALLENGES IN THE APPLICATION OF MACHINE LEARNING ALGORITHMS IN BIOMEDICAL RESEARCH

S. Marimuthu[1], Mani Thenmozhi[2], Melvin Joy[3], Malavika Babu[4] and L. Jeyaseelan[5*]

[1] *Department of Biostatistics, Christian Medical College, Vellore, Tamil Nadu, India.*
E-mail: marimuthu8421@gmail.com
[2] *Department of Biostatistics, Christian Medical College, Vellore, Tamil Nadu, India.*
E-mail: mani.thenmozhi@gmail.com
[3] *Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, UK.*
E-mail: melvinmj94@gmail.com
[4] *Centre for Trials Research, College of Biomedical and Life Science, Cardiff University.*
E-mail: malavikababu@gmail.com
[5] *College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE.*
E-mail: prof.ljey@gmail.com
Corresponding Author: E-mail: prof.ljey@gmail.com ; ljey@cmcvellore.ac.in

ARTICLEINFO

Received: 17 August 2022
Revised: 11 September 2022
Accepted: 25 October 2022
Online: 26 December 2022

To cite this paper:
S. Marimuthu, Mani Thenmozhi, Melvin Joy, Malavika Babu and L. Jeyaseelan. (2022). Challenges in the application of Machine Learning algorithms in Biomedical Research. *Journal of Applied Statistics & Machine Learning*. 1(2): pp. 117-126.

ABSTRACT

In recent years, the application of machine learning (ML) algorithms has increased rapidly in various domains. Extensively in assisting diagnosis and predicting the prognosis in health care research. However, the challenges in using these methods are less understood by the researchers. The aim of this article is to present the following challenges in using ML algorithms in biomedical research. The use of 'variable of importance' in the prediction as ML models do not provide coefficients or weights, relation to regression coefficients and predicting the diagnosis or prognosis of low prevalence (imbalance) diseases, and the adjustment to handle this imbalance using Synthetic Minority Over-sampling Technique called SMOTE, etc. Also, highlighted that the model selection with maximum accuracy or area under curve (AUC) statistics is alone not sufficient. The need for predictive values at various prevalence of outcome has to be highlighted. Simulation studies are recommended to evaluate the usefulness of SMOTE. The results of studies with the diseases prevalence 40% to 60% have to be used cautiously. Literature examples have been used to highlight the challenges.

Keywords: Challenges in ML, Imbalance data; Low prevalence; Machine Learning; SMOTE;

BACKGROUND

In recent years, the application of machine learning (ML) algorithms such as support vector machine (SVM), random forest (RF), Neural Networks, and Boosting methods has increased rapidly in various domains including health care, pharmaceutical, insurance, and stock market etc. Especially in health care, the goal of using these methods is to assist as a diagnostic tool to improve the accuracy (1), predicting admission at the emergency ward (2), accurate prognosis prediction (3,4), drug combination therapy (5) and etc. Deep learning algorithms such as convolutional neural networks (CNN), transfer learning are used to solve the problems related to image classification (6), image segmentation (7), predicting DNA sequence specificities (8) etc. The major reason behind these wide applications is its outstanding performance as compared to other classical models that are not assumption free. Many researchers have shown that the better accuracy or area under curve (AUC) in ML algorithms as compared to frequentist methods such as logistic regression (LR) and Cox proportional hazard model etc. They have compared these models with ML algorithm using Receiver Operating Characteristic (ROC) curves and reported the important variables etc. Based on these research articles, biomedical researchers are now using ML algorithms to identify variables of importance. Qin et al (2021) have aimed to construct ML models as an auxiliary diagnostic tools to improve the diagnostic accuracy of non-ST- elevation myocardial infarction (NSTEMI) (9). They have used three ML feature selection technique including RF and Extreme Gradient Boosting (XG Boost) and provided the list of variables which are important to predict the diagnosis. However, their findings did not help to predict whether a new subject will have NSTEMI or not without using computer with inbuilt algorithms. Though these ML models are useful in understanding the important variables, but they are not useful for prediction such as Trauma score or Acute Physiology and Chronic Health Evaluation (APACHE II) unless an infrastructure is developed to do this prediction (10–12). But this is possible in logistic regression and Cox models.

Luo *et al.* (2016) provided a set of guidelines on the use of ML predictive models within clinical settings (13). They have pointed out that the predictive models including RF and SVM have ability to make accurate predictions on unseen data as compared to the traditional statistical methods such as logistic regression. In order to optimize the prediction with large number of risk factors, often ML algorithms attempt to produce more difficult models. As a consequence, the researchers would not study the problem of overfitting in relation to the number of variables and also the prevalence of outcome.

In general, these algorithms perform well, however there are many challenges including a lack of transparency, replicability, ethics, and effectiveness (TREE) in the reporting and assessment of ML predictive models (14). Vollmer *et al.* (2019) proposed 20 critical questions to identify the pitfalls that can undermine ML/AI applications in health (14). Mani *et al.* (2021) have reported and raised a concern whether the ML algorithm suggests for risk factor analysis or to predict outcome (15).There are lots of challenges in using and presenting the results of ML algorithms. Many times it rises concern whether the research is meant for prediction or to find just the variables of importance. Therefore, the aim of this article is to present the challenges in using and reporting ML algorithms and also to present the suggestions.

METHODS

Imbalanced data: Data imbalance is a difference between the two classes of binary outcome (alive / dead). For example, in biomedical research, number of diseased patients is very less as compared to non-diseased patients. Invariably, the imbalance is due to prevalence of an event (diseased), which is usually lower and therefore the prevalence of non-diseased is far higher.

Over fitting: Finding the best fit to the training data causes a risk that the model will fit the noise in the data by memorizing various peculiarities of the training data rather than finding a general predictive rule. This is called "over-fitting" (16). Due to over fitting, the model performs too well on the training data but the performance drops significantly over the test data (17).

SMOTE: Synthetic Minority Over-sampling Technique (SMOTE) is a sampling technique which balances the subjects in the two classes (groups) by generating synthetic subject in minority class (18). Synthetic samples are generated as follows. Take the difference between the sample under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the sample under consideration. The process equals diseased and non diseased proportions and thus eliminates the bias that would arise due to imbalance in the group proportions (prevalence). That is, making the prevalence closer to 50%.

Validation: Validation technique is used to evaluate the performance of the trained model prediction by applying this model to the data which is not used for training. In prediction modelling, usually the model is trained with 70% data and validate with the remaining 30% data. As validity statistics sensitivity, specificity and predictive values are usually provided, besides Receiver Operating Characteristics (ROC) curve.

Black Box: The term "black box" is shorthand for models that are sufficiently complex that they are not straight forwardly interpretable to humans. This contrasts with models that are routinely used in medical research, such as linear and logistic regression, in which humans can refer to model coefficients to interpret the model and its predictions. Although not all ML algorithms are un-interpretable (more on this will follow), most ML algorithms producing state-of-the-art results including deep learning and ensemble models do suffer from this limitation (19).

Data: The following table of data has been used to illustrate the challenge due to small prevalence and its consequence in validity statistics. Let us assume that we have trained the data with 200 subjects, among them 40 (20%) subjects had disease.

Table 1: Hypothetical confusion Matrix of Actual and predicted outcomes : Trained data

		Truth		*Total*
		D+	*D-*	
Predicted class	**D+**	15 (37.5%)	10	25
	D-	25	150 (93.75%)	175
Total	40	160	200	

RESULTS

Impact of Imbalance (prevalence) on Model performance: The model was trained by maximizing metrics such as accuracy or AUC. In table 1, the accuracy was 82.5% (165/200) but the sensitivity was 37.5% (15/40). The model was not able to correctly classify the diseased patients as diseased, though the model performance is good. Here, the number of non-diseased cases dominates the model performance. In the goal of finding best hyper parameters (tuning), the overall performance is very good but not the sensitivity. Thus the challenge is that when the researchers use accuracy statistics to decide on the model, it is often not very useful for the clinical application. Especially in medicine, the sensitivity and specificity have to be better so that in the real time application (generalization) these statistics will be reasonably good.

Impact of prevalence on predictive values: The positive and negative predictive values are prevalence (pre-test probability) dependent statistics. Therefore, the researchers who decided to apply these findings in their settings need to know the prevalence of disease of their centre and the research article as well. In their settings, if the prevalence is different as compared to the reviewed paper, then they need to be cautious in using these findings. The prediction modelling concept is used in Diagnostic test research. The guidelines for critiquing for diagnostic test is very much applicable in ML research as well. But these principles are often ignored in ML research. Example, the overall Treatment related mortality (TRM) in multicentre haematological cancer study is 20%. There are ten centres involved in the study. If one centre has the mortality of 10% then can we use the predictive values derived from the overall morality of 20%? The question remained to be answered is, how do we adjust the predictive values for varying prevalence? Therefore, most of the ML research identify the variables of importance for prediction. However, they do not provide a model for prediction or classification such as Glasgow Coma Scale.

Qin *et al* (2021) have evaluated six algorithms in order to select the best one for the above research question to construct ML models as auxiliary diagnostic tools to improve the diagnostic accuracy of non-ST-elevation myocardial infarction (NSTEMI) (9). At the end of good analyses, they were able to suggest a ML algorithm XG Boost as the best as compared to other algorithms and Logistic regression model as well. The retrospective data of 1409 patients with NSTEMI and 1469 patients with unstable angina pectoris was used for the analysis. Thus the percent of NSTEMI in the study was 48.9%. The diseased to non-diseased ratio was about 50:50, which is well balanced. Therefore, in such situations the accuracy and the validity statistics are expected to be higher. However, the question to be asked is, what is the prevalence of NSTEMI in a given hospital? Will the variables of importance and the ML model be same if the prevalence of NSTEMI is 10%? These findings cannot be used in real life situation. Therefore, the results have to be used with caution from this paper.

Impact of SMOTE: As presented above, Qin et al (2021) have identified the variables of importance to improve the diagnostic accuracy of non-ST-elevation myocardial infarction, based on the three ML algorithms (9). The NSTEMI and no NSTEMI ratio was about 50% and 50%. It is very unlikely to get such diseased and non-diseased case ratio in a cohort study or the cases were selected in such a way that the ratio is nearly matched. The algorithm is expected to provide higher accuracy with such ratio. However, they did not provide the sampling scheme. Therefore, the results (variables of importance) of this paper have to be used cautiously. Again this revolves around the prevalence or pre-test probability of disease. Such ratio is very uncommon unless someone uses SMOTE method to optimize the disease and non-disease ratio. The users of ML algorithm related papers are expected to be aware of these challenges.

Absence of Regression coefficients (Black Box): Regression coefficients play an important role in parametric approach such as logistic regression and Cox Proportional Hazard model etc. These coefficients helped us to develop and use diagnostic and prognostic tools such as Trauma score (10,12). This scale is used to objectively describe the extent of impaired consciousness in all types of acute medical and trauma patients. The scale assesses patients according to three aspects of responsiveness: eye-opening, motor, and verbal responses. The weights for these three aspects were derived objectively using multivariable LR methods and evaluated in various scenarios. But this is not possible in ML algorithm. Algorithms such as RF and XG boost are incorporating the concept of 'important variable,' which help us to identify the important variables associated with the outcome, but they do not provide regression coefficients for these variables. Therefore, the ML algorithms may not be useful for risk factor analysis. If someone intends to predict the disease outcome, then the researchers need to set up real time predictive algorithm. This needs real time computing with the algorithms. However, this is a challenge in multicentre studies whether the prevalence of outcome of interest varies from site to site.

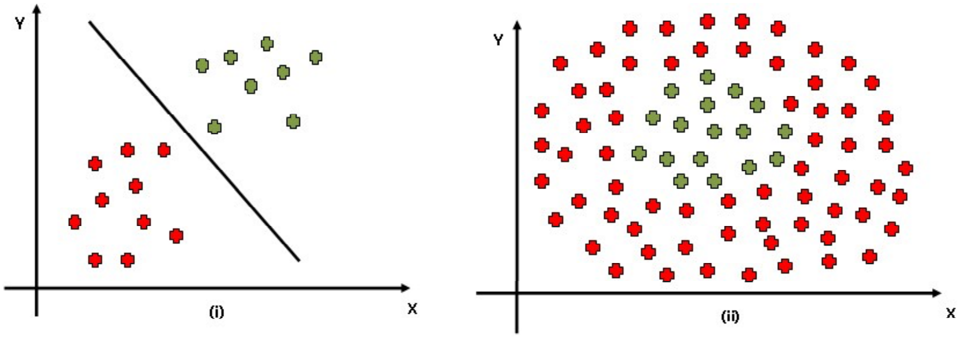

Figure 1: Simple model Vs complex model

In figure 1, obviously (i) is a simple problem, we can classify the points by using logistic regression and interpret the relation between the explanatory and outcome variable using the regression coefficients. Whereas, (ii) is more complex than (i), therefore we need a complex algorithm to classify the points perfectly. In this case, SVM with RBF kernel or ANN may be useful. However, we cannot get coefficients like logistic regression and it is not possible to interpret the relationship between explanatory and outcome.

Consistency: While dealing with ML algorithms, the consistency or stability of the model performance is essential. For example, in the CART method the trees are grown by splitting the variables. Bias and variance of such a model is expected to be high. Yoon *et al.* (2020) mentioned that the bootstrap aggregation by re-sampled cases helps to achieve stability, consistency of the prediction and avoid over-fitting thus increase the performance (20). As number of trees increases the variance will decrease. RF, boosting algorithms such as Gradient Boosting, AdaBoost, and XG Boost and the stacked ensemble models (combine various models) use bootstrap algorithms to improve consistency as well as accuracy. Therefore, the suggestion is to incorporate the bootstrap methods while executing the ML algorithm. Therefore, it would be ideal to incorporate bootstrap method in certain algorithms such as CART and SVM.

Over-fitting: Due to the over-fitting, performance of the model in trained data is much higher than test data. Presence of noise, the limited size of training data, complex classifier, and including too many variables in the models causes over-fitting (13,17). Including too many variables in the model increases the chance of noise and inaccurate data. Therefore, more complex algorithm needs to classify the subjects correctly. Choose the most important features (variables) from the set of variables then fit a model rather than using all the variables. Lasso regression or RF or other ML feature selection methods can be useful to select the important features.

DISCUSSION

Recent evaluation of artificial intelligence (AI)–based facial Recognition software has renewed concerns about the inadvertent effects of AI on social bias and inequity. Academic and government officials have raised concerns over racial and gender bias in several AI-based technologies, including internet search engines and algorithms to predict risk of criminal behaviour. Companies like IBM and Microsoft have made public commitments to "de-bias" their technologies, whereas Amazon mounted a public campaign criticizing such research. As AI applications gain traction in medicine, clinicians and health system leaders have raised similar concerns over automating and propagating existing biases (21,22). Thus it is natural or expected fact that the models developed to be biased to some extent. The end users of these models to be aware of the challenges that would cause these concerns before they use models. This article is an attempt to provide such challenges in ML algorithm-based models.

Black Box: The recent developments in ML algorithms have made these modelling useful. However, there is a misunderstanding that the researchers think that these algorithms could be used to predict whether the patient's prognosis is going to be better or not. Unfortunately, this is easily available, unless the research team provides a real time algorithm and infrastructure to predict someone's prognosis. Mostly the authors present the variables of importance which would behave like risk factors with limited use. These may not be able to provide statistics such as odds ratio or relative risk. As the algorithms are very intensive and work inside the black box, the regression coefficients are not available. However, each subject's probability of getting an event is provided.

Causal Reasoning: Pearl (2018) has reported that ML algorithms operate, almost exclusively, in a statistical or model-free mode, which entails severe theoretical limits on their power and performance (23). Such systems cannot reason about interventions and retrospection and, therefore, cannot serve as the basis for strong AI. He also suggested that the ML algorithms need the guidance of a model of reality, similar to the ones used in causal inference tasks.

SMOTE: The method of handling imbalance in the prevalence disease using SMOTE has been good. However, the Engineering community who developed this method or the end users of this method need to understand the implications in the predictive values. This is less understood by the end users in the medical community. A simulation study has to be done to study the implications of SMOTE method in terms of predictive values and suggest the use of Bayesian methods to correct the bias.

Limitations: This article provided the bird's eye view of challenges in ML algorithms. In order to understand thoroughly the impact of SMOTE in prediction simulation studies need to be done. There are many articles which have misused the concept of sampling and never reported the sample size calculations etc. But these articles have not been referred here.

Reference

1. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J, *et al.* Single Reading with Computer-Aided Detection for Screening Mammography. *New England Journal of Medicine*. 2008 Oct 16;359(16):1675–84.
2. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLOS ONE. 2018 Jul 20;13(7):e0201016.
3. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, *et al.* Scalable and accurate deep learning with electronic health records. npj Digital Med. 2018 May 8;1(1):1–10.
4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J.* 2015;13:8–17.
5. Tsigelny IF. Artificial intelligence in drug combination therapy. *Brief Bioinform.* 2019 Jul 19; 20(4):1434–48.

6. Lee C, Jang J, Lee S, Kim YS, Jo HJ, Kim Y. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. *Sci Rep.* 2020 Aug 13;10(1):13694.

7. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med Image Comput Comput Assist Interv.* 2013;16(Pt 2):246–53.

8. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol.* 2015 Aug 1;33(8):831–8.

9. Qin L, Qi Q, Aikeliyaer A, Hou WQ, Zuo CX, Ma X. Machine learning algorithm can provide assistance for the diagnosis of non-ST-segment elevation myocardial infarction. *Postgraduate Medical Journal [Internet]. 2022 Feb 16 [cited 2022 Jun 1]; Available from: https://pmj.bmj.com/content/early/2022/02/15/postgradmedj-2021-141329*

10. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. *J Trauma.* 1989 May;29(5):623–9.

11. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med.* 1985 Oct;13(10):818–29.

12. Jeong JH, Park YJ, Kim DH, Kim TY, Kang C, Lee SH, *et al.* The new trauma score (NTS): a modification of the revised trauma score for better trauma mortality prediction. *BMC Surg.* 2017 Jul 3;17(1):77.

13. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, *et al.* Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016 Dec 16;18(12):e323.

14. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ.* 2020 Mar 20;368:l6927.

15. Thenmozhi M, Marimuthu S, Jeyaseelan L. Utility of ML algorithms: Are we predicting the outcome or finding the risk factors. *Postgraduate Medical Journal [Internet]. 2022 Mar 30 [cited 2022 Jun 3]; Available from: https://pmj.bmj.com/content/early/2022/02/15/postgradmedj-2021-141329.responses*

16. Dietterich T. Overfitting and undercomputing in machine learning. ACM Comput Surv. 1995 Sep 1; 27(3): 326–7.

17. Ying X. An Overview of Overfitting and its Solutions. *J Phys: Conf Ser.* 2019 Feb; 1168: 022022.

18. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research.* 2002 Jun 1;16:321–57.

19. Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Canadian Journal of Cardiology.* 2022 Feb 1; 38(2):204–13.

20. Yoon HJ, Klasky HB, Gounley JP, Alawad M, Gao S, Durbin EB, *et al.* Accelerated training of bootstrap aggregation-based deep information extraction systems from cancer pathology reports. *Journal of Biomedical Informatics.* 2020 Oct 1; 110: 103564.

21. Gianfrancesco M, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA internal medicine.* 2018;
22. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA.* 2019 Dec 24; 322(24): 2377–8.
23. Pearl J. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. WSDM. 2018.