JCBI
Journal of Computing & Biomedical Informatics ICASET
ISSN: 2710 - 1606 2024

Research Article
Collection: Intelligent Computing of Applied Sciences and Emerging Trends

Leveraging Machine Learning for Advancements in Epidemiology and Health Outcomes Research

Madiha Ashraf[1*], M. Asim Rajwana[1], Abdul Majid Soomro[1], Qudsia Zafar[1], and Muhammad Akhter[1]

[1]National College of Business and Economics, Lahore (Multan Campus), 66000, Pakistan.
[*]Corresponding Author: Madiha Ashraf. Email: madihaashraf686@gmail.com

Academic Editor: Salman Qadri **Published:** April 01, 2024

Abstract: Machine learning modelings that involve use of epidemiologic data to predict disease outbreaks are beginning to appear in professional studies. Such techniques spark hope in patients and doctors alike, they show us what can be done and what needs to be done as never happened before. Here is a tutorial on how to create supervised machine learning models with corresponding literature resources. Source: From picking a suitable sample, delineating the features through training, test, and evaluation of performance, machine learning by the end-to-end approach can be a really stressful thing. In the article, we take the reader by the through steps in the process and explore groundbreaking concepts on machine learning, such as treatment effects and explaining machine learning models' output.

Keywords: Machine Learning; Epidemiology; Health Outcomes Research.

1. Introduction

The term epidemiology refers to the continuous track tests of where disease comes from and spreads. The population health situation and an amazing survival ability of humans can be explained provided there is evidence gained from conducting epidemiological study [1]. Data analytics are core to the epidemiology whose capabilities have skyrocketed over recent years thanks to the improvements in the computing technologies. With the variety of disciplines that make up the sector, as well as the diverse nature of the targeted populations, this field plays host to the highest incidence of novel modelling methods and techniques in the published literature[2]. So far, Frequentist statistical methods are the main tools employed for analysis in epidemiological studies at the present. These ways are normally confusing for the nonprofessionals and they are soaked in the construction of the hypothesis and the resulting probabilities that are either supporting or disapproving those hypotheses [3]. The basic statistical tests such as t-test and multiple regression modeling are popular methods for testing the hypothesis of defining the associations or treatment effects between the predictor and the target/outcome variables [4]. These traditional statistical approaches are referred to as the data culture and which you can call the "statistical data culture". A regression-type model that is often employed in epidemiology to foretell a dependent variable using multiple independent variables can be classified by the purpose of a model or model building [5]. Parametric methods (regression analysis) focus on generating estimates of therapeutic effect or the size and the strength of association of independent variables with the dependent variable (correlation analysis) and explanatory models [6]. The strategies of modeling are of great help to investigators and practitioners since they provide useful information. A majority of traditional modeling methods are data-centric, hence they take the data into account while formulating numerous assumptions [7]. Epidemiologists have made scientific conclusions such as linearity, no multicollinearity, and equal risk/odds/hazards throughout time which should be taken for granted. With the increased data available for analytics, one proving the point of "curse of dimensionality" how Richard Bellman put it [8]. Here, research questions become more complicated than usual, standard modeling assumptions are difficult to meet with nonlinearity is ubiquitous, and new approaches are necessary. Innovation in artificial

intelligence has been a last-mile effort in medicine, although it has not been as extensively deployed in the population health realm over the past several years [9]. The aim of this article is to provide historical evidence with respect to the machine learning algorithms on epidemiological and health outcomes research, giving special attention to the supervised learning methods.

"Machine learning" is the generic term utilized for different types of algorithms and approaches which relies heavily on mathematical modeling. The regression that we are concerned with in this paper being the regression analysis is however, entirely different from the 'regression' in epidemiology, the latter being a wide variety of frequentist regression models such as logistic, linear, and Cox proportional hazards which often are the ones used in epidemiology and biostatistics[10]. From the very notion of machine learning which is dated back to the early 50s come the possibilities for the computer to imitate the brain processes of human beings such as pattern recognition, matching, and decision-making. This work was followed by the research of Arthur Samuel, which developed the first computer program playing checkers against people, and Frank Rosenblatt, who developed a neural network model which mimics the neuronal processes to solve computing tasks. This breakthrough has given rise to machine learning algorithms that are capable of solving a wider range of problems in learning contexts. These algorithms are generally categorized into two groups i. e. supervised and unsupervised. Whereas supervised models tend to be utilized for predicting outcome (essentially as the case with regression models), the unsupervised models are more categorized as descriptive, rather than predictive in nature. Nevertheless, the unlabeled models are preferred when the goal is to identify hidden patterns not related to a specific label. Hereafter we proceed with the supervised models in the epidemiology review. Despite the fact that the word viz., machine learning has spread in all corners today, "statistical learning" word also comes across in the literature. This disparity in terminology is owing to the novel techniques that consist of the two popular methods: frequentist biostatistical procedures (i.e. hypothesis testing's) and algorithmic approaches which are generally applied to machine learning models. It gives rise to "statistical learning", a contested field which is located between classical biostatistics and machine learning, thus online in-between the two. Nevertheless, the terms used in epidemiology and machine learning may have minor differences despite occasionally being used to describe similar concepts. In this review, we adopt machine learning terminology to our effect.

2. Literature Review

The literature contains different methods that are used to discover the appropriate sample sizes for machine learning models. On the contrary, the sample sizes calculations are hard to make as the machine learning models rely on algorithms mostly. They do not involve frequency measures such as p-value that are used to calculate sample size in the classical statistical inference nor do they calculate effect size statistics that are also important [11]. In case of unsupervised models, the size of the sample should reflect the research focus and the variance expected on the data considered. Hence, due to sample, size computations might not be needed for hypothesis generation or data reduction since unsupervised models. For example, small datasets have been efficiently used to not only find inflammation markers in hospitalized patients with pneumonia but also to classify them according to their similarities or dissimilarities. In brief, there are multiple sorts of suggestions and guidelines for machine learning sample size estimation you can choose from [12]. Machine learning sample size estimations in publications are often tailored to a particular area of specialization. For example, genetic epidemiology might work with lower data sets because of the privacy measures taken. g. In addition to that, the small sample size (<100 rows) will be in contrast with a moderate size (several hundred cases) required for the evaluation of behavioral/ cognitive domain using functional magnetic resonance imaging (fMRI). Nevertheless, it should be highlighted that the best number of sample points is decided according to the data and as such a good size and quality of features [13]. Points are meaningless if they contradict features that represent labels, however many features the model has. Additionally, encompassing too many features and sparse instances in comparison to the labels would result in a situation where the model would be incapable of correlating patterns with the labels for the whole feature spaces and the model would not operate well in production [14]. Similarly, in data sampling a certain result can be obtained irrespective of the size of data. In machine learning, the total number of elements used in the sample must be examined properly along with the anticipated results to be accurate and generalizable that may be far larger than initially expected.

Parsimony plays a pivotal role in epidemiology as a crucial tenet for avoiding overfitting [15]. Data dimensionality necessitates adequate explanatory variables and cherry-picking appropriate predictors to be effective. Overfitting is another major issue in machine learning model just as it occurs for regression; to avoid this model would need to have parsimony which is related to feature selection in which the features of the data sets are deliberately chosen [16]. This technique is particularly significant for machine learning models because they are wide-ranging and tend to tap into data sets that were initially gathered for purposes unrelated to a certain hypothesis, as is the case with the data sets of genetic epidemiology or even electronic medical records. The like of data sets usually hold a vast amount of features yet less cater towards the construction of the required model [17]. The method described selects and limits the number of variables to be used for model building (called a feature set) in order to discover antigens that can be used for vaccinology. This method can be applied to any data set (e. g. 'omics) data set. Choosing a set of parsimonious features to suit a particular study outcome is comparably more complex than the general impact on the outcomes of the study. Alongside the aforementioned, we encounter a couple of additional reasons to delete some features from one's training data set prior to feeding that data into a machine learning model. A simulation model will be trained faster, which also attractive when complex modeling and ubiquitous local computation is a focus point instead of cluster computing. The second technique is that lessening the scope of redundant features or features which do not affect the outcome might bring along the probability of model overfitting [18].

The choice of feature selection can also be made in multiple ways, such as through selection of clinically relevant features, simple bivariate relation between features, and feature importance scores. Along with LASSO regression, another machine learning model called least absolute shrinkage and selection operator (LASSO) may be suitable too for feature extraction. The application of genetic algorithms for the purpose of feature selection has grown wide with a common objective of studying the impact that uncontrolled comorbidities have on clinical outcomes experienced by patients with pneumonia admitted into hospitals. In any case, investigators have proposed that the precision and stability of the model should be the next indication that should be considered when using feature selection algorithms [19]. Ultimately, these models run the risk of overfitting in such cases. With this 'omics era in which many data points are already available for epidemiologists to analyze, the feature set has been bulked up a lot. A veritable range of different methods for selecting features have been suggested in literature. Instances of ranked guided iterative feature elimination (RGIFE) containing promising characteristics of a clinically relevant biomarker are a case in point. Even so, in a more complex manner similar to explanatory regression model development, pure automation of feature selection is very probably not a suitable approach alone. In almost all areas, the experts from the domains are to be consulted for wise choices of the features to be included for the models that will be built [20].

3. Materials and Methods

The growing discipline of healthcare research, which is continuously attentive to fresh and improved strategies, is geared towards to comprehend the nature of disease occurrence and patient outcomes. This research focuses on the ability of machine learning (ML) to be truly revolutionary for both epidemiology and health outcomes. Through the use of huge data sets and robust algorithms, ML presents prospect of being the instrument of the immense transformation of the knowledge about disease causes, risk factors and progression.

The techniques of machine learning would be integrated with the standard methods of epidemiological and health outcomes research to make a study reliant on this approach. Data collection is the crucial initial step which is going to involve those who are patients' data, environmental factors and DNA type information. This rich data will be carefully cleaned and processed for machine learning purposes afterwards.

After that, the data can be used to train different machine learning models. Using these frameworks, the taken models could be supervised learning for tasks such as disease prediction and classification or finding hidden patterns in complex datasets using unsupervised approaches. This will happen through the repeated training of models by constantly improving their performance, finally targeting high accuracy and generalizability.

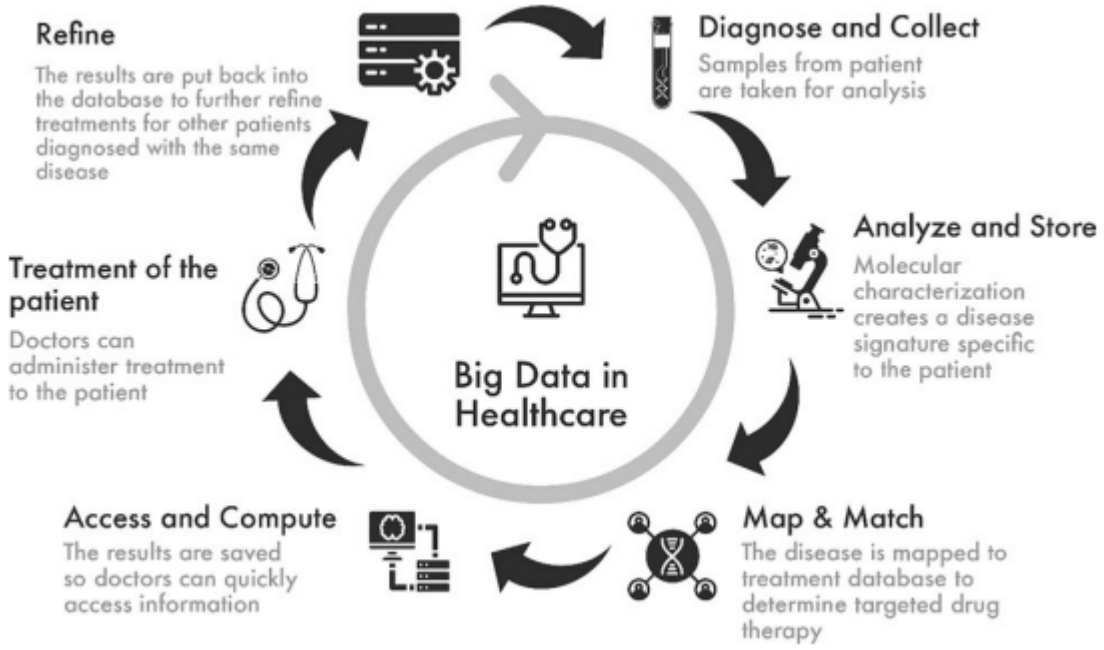

Figure 1. Big Data Analysis for Treatment [21]

The next step is to train, validate and finally to apply these models to the real world epidemiological and health outcomes questions. For example, determining the risk factors for particular diseases, predicting the spread of outbreaks or evaluating the correctness of various treatment strategies. This study will utilize the capabilities of machine learning to unveil innovative evidences that may be the key to the development of much more effective public health interventions and personalized healthcare approaches.

4. Conclusions

In conclusion and this analyzation has shown that the role of machine learning (ML) as the engine of molecular epidemiology and of health outcome research is twofold. Through the use of big data and automated algorithms, Ml unique a different natural on disease etiology risk factors and progression process. The discoveries of such research lead to far lasting ramifications of different beneficiaries of health fields such as individuals, physicians, health institutions, and the entire health industry. Epidemiologists will be able to implement ML algorithms, which will ultimately help to identify unrecognized disease clusters, create more accurate predictive models, and trace the environmental or social factors determining the onset of a particular disease. Data extracted from genomic research can be used by public health departments to devise specific intervention and funding models and, consequently, enrich health measures of disease prevention and control.

For healthcare providers, the ML (machine learning) tools can open up the doors for medicine fostering an individual approach. ML models, having such capabilities as scanning patient data integrated with medical history, genetic information and lifestyle factors, anticipate disease risks and suggest personalized preventive measures to individual patients. Furthermore, this models help in treatment plan optimization: treatment that is individually targets for each patient besides increasing possibility of successful outcomes. Besides effects on patients alone there are more general impacts, which are capable of improving the nature and types of health service provision. Problems in healthcare resource allocation can be minimized through ML, as its ability to predict resource demands through disease trend analysis allows institutions to timely adjust resource allocation and streamline unnecessary staffing, thereby ensuring efficient and cost-effective care. Moreover, machine learning tools can automate some administrative processes to equalize the information flow, making the healthcare system more efficient.

References

1. Adenyi, A.O., et al., Leveraging big data and analytics for enhanced public health decision-making: A global review. GSC Advanced Research and Reviews, 2024. 18(2): p. 450-456.
2. Deng, X., S. Cao, and A.L. Horn, Emerging applications of machine learning in food safety. Annual Review of Food Science and Technology, 2021. 12: p. 513-538.
3. Ebulue, C.C., et al., Leveraging machine learning for vaccine distribution in resource-limited settings: A synthesis of approaches. International Medical Science Research Journal, 2024. 4(5): p. 544-557.
4. El Hechi, M.W., et al., Leveraging interpretable machine learning algorithms to predict postoperative patient outcomes on mobile devices. Surgery, 2021. 169(4): p. 750-754.
5. Flores, A.M., et al., Leveraging machine learning and artificial intelligence to improve peripheral artery disease detection, treatment, and outcomes. Circulation research, 2021. 128(12): p. 1833-1850.
6. Glymour, M.M. and K. Bibbins-Domingo, The future of observational epidemiology: improving data and design to align with population health. American Journal of Epidemiology, 2019. 188(5): p. 836-839.
7. Gupta, A., et al., Advancement in deep learning methods for diagnosis and prognosis of cervical cancer. Current Genomics, 2022. 23(4): p. 234.
8. Hederman, A.P. and M.E. Ackerman, Leveraging deep learning to improve vaccine design. Trends in Immunology, 2023.
9. Ijeh, S., et al., Predictive modeling for disease outbreaks: a review of data sources and accuracy. International Medical Science Research Journal, 2024. 4(4): p. 406-419.
10. Kino, S., et al., A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. SSM-population Health, 2021. 15: p. 100836.
11. Lee, W., et al., A scoping review of the use of machine learning in health economics and outcomes research: part 2—data from nonwearables. Value in Health, 2022. 25(12): p. 2053-2061.
12. Masud, M., et al., Leveraging deep learning techniques for malaria parasite detection using mobile application. Wireless Communications and Mobile Computing, 2020. 2020: p. 1-15.
13. Mhasawade, V., Y. Zhao, and R. Chunara, Machine learning and algorithmic fairness in public and population health. Nature Machine Intelligence, 2021. 3(8): p. 659-666.
14. Mir, M.H., et al., IoT-enabled framework for early detection and prediction of COVID-19 suspects by leveraging machine learning in cloud. Journal of healthcare engineering, 2022. 2022.
15. Morgenstern, J.D., et al., Perspective: Big data and machine learning could help advance nutritional epidemiology. Advances in Nutrition, 2021. 12(3): p. 621-631.
16. Morrow, A.S., et al., Leveraging machine learning to identify predictors of receiving psychosocial treatment for Attention Deficit/Hyperactivity Disorder. Administration and Policy in Mental Health and Mental Health Services Research, 2020. 47(5): p. 680-692.
17. Padhi, A., et al., Transforming clinical virology with AI, machine learning and deep learning: a comprehensive review and outlook. VirusDisease, 2023. 34(3): p. 345-355.
18. Rajyalakshmi, P., et al. Leveraging Big Data and Machine Learning in Healthcare Systems for Disease Diagnosis. in 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM). 2022. IEEE.
19. Reddy, M., R. Naveed, and T. Shah, Urban Health Planning in the Age of AI: Advancements and Opportunities in Machine Learning. International Journal of Sustainable Infrastructure for Cities and Societies, 2023. 8(1): p. 38-52.
20. Wiens, J. and E.S. Shenoy, Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical infectious diseases, 2018. 66(1): p. 149-153.
21. Subrahmanya, S.V.G., et al., The role of data science in healthcare advancements: applications, benefits, and future prospects. Irish Journal of Medical Science (1971-), 2022. 191(4): p. 1473-1483.