Review Article
Volume 9:2, 2021

Journal of Health Education Research & Development

ISSN: 2380-5439 Open Access

Machine Learning in Public Health: A Review of the Problems and Challenges

MD Asadullah*, **Mamunar Rashid, Priyanka Basu, Md Murad Hossain**

Department of Health Education Bangabandhu Sheikh Mujibur Rahamn Science and Technology University, Gopalganj, Bangladesh

Abstract

In recent years Machine learning that has been used for disease diagnosis and prediction in public healthcare sector. It plays an essential role in healthcare and is rapidly being applied to education. It is one of the driving forces in science and technology, but the emergence of big data involves paradigm shifts in the implementation of machine learning techniques from traditional methods. Computers are now well equipped to diagnose many health issues with the availability of large health care datasets and progressions in machine learning techniques. Several machine learning techniques have been used by researchers in public health. Several of these methods, including Support Vector Machines (SVM), Decision Trees (DT), Naïve Bayes (NB), Random Forest (RF) and K-Nearest Neighbors (KNN), are widely used in predictive model design research, resulting in effective and accurate decision-making. The predictive models discussed here are based on different supervised ML techniques as well as various input characteristics and data samples. Therefore, the predictive models can be used to support healthcare professionals and patients globally to improve public health as well as global health. Finally we provide some basic problems and challenges which face the researcher in public health.

Keywords

Machine learning • Prediction • Classification • Public Health • Disease

Introduction

Machine learning, a method of developing a prototype that learns to enhance its quality through experience, belongs to the context of artificial intelligence and is increasingly being used in various fields of science [1]. Such algorithms can be applied to help track the progress of a person, what variables make their symptoms worse, predict how long they would take to be completely rehabilitated, etc. It is likely to deliver technically superior results, but it is not going to be perfect. As such, while machine learning can deliver superior technical performance, inequities can be compounded. The intervention was particularly effective among the group with a moderate likelihood of participation. Targeting using the results of the prediction model using the machine-learning method has been useful in identifying suitable intervention targets. Traditional machine-learning approaches have been successful, mostly because the complexity of molecular interactions has been reduced by investigating only one or two dimensions of the molecular structure in the feature descriptors. A number of different ML classifiers are experimentally validated into a real data set in the present study [5]. Machine learning is involved in many of these, but streaming data is only addressed in a few plays. The machine learning library consists of common learning algorithms such as classification, clustering, collaborative sorting, etc., which is useful when dealing with problems with machine learning. Machine learning typically extends these methods to cope with high dimensionality and nonlinearity, which in wearable sensor data is of particular importance. It overlaps with artificial intelligence, but traditional biomedical statistics usually recognize the problems it seeks to solve. Extraction of the function renders the issue of machine-learning traceable because it greatly reduces the number of data dimensions. These techniques can help enhance the ability to discriminate by combining multiple metabolites ' predictive abilities. However, these methods are monitored and therefore various validations are key factors in preventing over fitting. In this paper, a new approach is proposed to automatically identify fundus objects. The method uses pre-processing techniques for

***Address to correspondence:** MD Asadullah, Department of Health Education Bangabandhu Sheikh Mujibur Rahamn Science and Technology University, Gopalganj, Bangladesh; Email: asadullahstat@gmail.com*

Received: 08 September, 2021; **Accepted:** 22 September, 2021; **Published:** 29 September, 2021

image and data to improve the performance of classifiers for machine learning. Machine learning techniques are applied to these data, which are useful for data analysis and are used in specific fields. Recently it can be used to analyze medical data and are useful for medical diagnosis to identify various complex diagnostic problems. We can improve the accuracy, speed, reliability and performance of the diagnosis on the current system by using machine learning classification algorithms for any particular disease. It has been used to estimate vegetation parameters and to detect disease, with less consideration being given to the effects of disease symptoms on their performance.

Machine Learning in Public Health

Machine learning plays an essential role in the healthcare field and is rapidly being applied to healthcare, including segmentation of medical images, authentication of images, fusion of multimodal images, computer-aided diagnosis, image-guided therapy, image classification and retrieval of image databases, where failure could be fatal. Statistical models developed using machine-learning methods can be viewed in many ways as extensions from epidemiology and health econometrics of more conventional health services research methodologies. In view of the wide availability of free packages to support this work, many researchers have been encouraged to apply deep learning to any data mining and pattern recognition topic related to health informatics. In medical fields, machine learning has also shown promise when the aim is to discover clusters in the data such as therapeutic choice imaging research. Here, the new features can be checked with a radiologist or neurologist expert assessment which varies from the prediction environment where observed marks exist in the data. Screening and prognosis of patients with cancer use methods for pattern recognition and identification such as machine learning. The repository should highlight the specifications of clinical machine learning tasks and thus motivate the ML community by providing a platform for the publication, exchange / collection of data sets, benchmarking of statistical evaluators and methods for challenging machine learning problems. The main purpose of applying the classification method is to allow healthcare organizations to provide accurate medication quantities. At every stage of development and application of machine learning in advancing health, ethical design thinking is essential. To this end, honesty and innovation physicians will work closely with software and data scientists to re-imagine clinical medicine and foresee its ethical implications. It is crucial that data from mobile health and consumer-facing technologies be systematically validated, especially in cases where dynamic intervention is provided [2]. Three developments in machine learning may be of interest to public health researchers and practitioners. Machine Learning techniques have showed success in prediction and diagnosis of numerous critical diseases. Some set of features are used in this strategy to represent each instance in any dataset. Research comparing the quality of different classification /prediction methods to predict the existence of disease, disease etiology, or disease subtype is minimal. For many types of medical diagnosis, a good machine learning approach to classification will apply.

Challenges in Public Health

Overall, health systems face multiple challenges: rising disease burden, multimorbidity and disability driven by aging and epidemiological transition, increased demand for health services, higher social expectations, and increased health spending. Healthcare offers unique machine learning challenges where the requirements for explaining ability, model fidelity, and performance in general are much higher than in most other fields. Ethical, legal and regulatory challenges are unique to health care since health care decisions can have an immediate impact on a person's well-being or even life. The primary focus in health informatics is on computational aspects of big data, which includes challenges, current Big Data Mining techniques, strengths and limitations of current works, and an outline of directions for future work. A major challenge is posed by the high volume of healthcare data, the need for flexible processing and support for decentralized queries across multiple data sources. Global health as an approach to the current situation and challenges, and the use of digital health as an ideal way to address health challenges associated with conflict-affected environments. There are a number of ways in which the proposed models of machine learning can help address public health challenges. The regularity, reliability and granularity of available data is a major challenge in tracking population health. Model estimates can play an important role in strategic decision-making if they are able to achieve sufficient precision, and machine learning models can provide a route to this required level of precision. Several writers describe different challenges in public health

Challenges	Description
Development	Challenges in the acquisition of talent and growth capital
Data schema	Increasing the burden of disease, multimorbidity and disability driven by aging and epidemiological transition
Ethics, laws and regulations	Health care choices can have an immediate impact on a person's well-being or even life.
Epidemic	Social health inequalities, a small number of local healthcare professionals, and a weak infrastructure for healthcare.
Big Data	Data mining methods, advantages and weaknesses of current works and recommendations for future work
Treatment effect	Treatment of patient outcomes in order to select the correct treatment
Clinical Data	Real clinical information environment, incomplete and erroneous data.
Data regularity, timing and reliability	The regularity, pacing and granularity of available data is the control of population health.
Characteristics identifying	The features of communities, ecosystems and policies are defined in population health
Health Tackling	Health as an approach to the existing situation and challenges.
Dataset imbalance	Forming an ensemble of multiple models with matched numbers of

	positive and negative slides trained on data subsets.
Biomarkers identify	Build diagnostic, prognostic or guided therapy predictive models
Screening	The area of early detection of cancer is packed with highlighting cautionary tales.

Table 1: Public health Challenges.

Problem Statement

In public health, reducing constraints such as lack of resources (human and logistic) in healthcare centers, high population dispersion and lack of infrastructure. One problem with the concept of "data health" is the lack of a practical idea of effective and efficient implementation of healthcare programs: each insurer has sought effective strategies through trials and errors [4]. The main problem is the unstructured of the medical reports. High complexity and noise issues result from the multisource and multimodal nature of healthcare data. Additionally, the high-volume data also has problems with impurity and missing values. All these issues are difficult to handle in terms of both size and reliability, although a range of methods have been developed to improve data accuracy and usability [2]. Machine learning methods are the leading option for achieving a better result in classification and prediction problems. In a wide range of machine learning (ML) problems, classification plays a major role. Another major issue with the collection of data is the potential lack of label accuracy. Over fitting is a potential problem in machine learning. A general problem is that several of the existing datasets are difficult to use in terms of permission. Table 2 displays the numerous public health issues facing them.

Problem	Description
Classification	The situation was linear in nature for all armed and unarmed group datasets
Scalability	Exists with two of the most widely used interpretable machine learning models
Lack of infrastructure	Lack of resources in health care centers (human and logistic), high population dispersion
Effective and Efficient	Through trial and error, every insurer tried effective strategies
Exchange health information securely	Scientists and clinicians across institutional, provincial, or even national jurisdictional boundaries across a given healthcare organization.
Over fitting	Because of its storage limitation, it may not be appropriate for very large datasets with high dimensional features
Data Imbalanced	Which are commonly used to resolve big data clinical databases.
Clinical unstructured notes	The multisource and multimodality of health care data leads to high complexity and noise problems
Impurity and missing	The high-volume data also has problems with impurity and missing values
Missing variables	This results in the normal multivariate methods, while machine-learning approaches can still be appealing for other reasons
Prediction	The computer is equipped with a set of data to improve the classification model after it can be used for future predictions

Table 2: Problem Statement in Public Health.

Dataset

To generate the most effective results, machine learning algorithms are used to analyze data over and over again. Machine learning currently provides the basic machine for scrutinizing imaginative information. Today, medical clinics are very well equipped with fully automatic machines, and these machines produce tremendous amounts of data, then collect and exchange these data with information systems or doctors to take the necessary steps. Machine learning methods can be used to examine medical data and various technical diagnostic conditions can be found in medical diagnosis. Using machine learning, systems take patient data as an input such as symptoms, laboratory data and some of the important attributes and produce reliable diagnostic results. Depending on the reliability of the test, the computer must determine the information for the future reference will be used as learning and qualified dataset. Different Authors are used to different data determine the quality of the proposed classifiers which display.

Classification Technique

In many real-world issues, classification is one of the most important decision-making techniques. The higher number of samples selected for many classification problems, but this does not lead to higher classification accuracy. Supervised machine-learning algorithms are mainly used for classification or regression issues where the patient sample class label is already available. Classification tasks are found in a wide range of decision-making tasks in various fields such as medicine, science, industry, etc. Several approaches are suggested in the literature on how to solve classification problems [3]. In medical context, the identification quality of commonly used machine learning models, including k-Nearest Neighbors, Nave Bayes, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression. In this research paper we conclude various research papers in a tabular form (Table-4) showing different methodologies and compare accuracy

Technique	Disease Name	Highest Accuracy
SVM,RF,KNN,DT	Parkinson's	SVM=97.22%
NB,KNN,C4,5DT,RF,SV M	Liver Disease	KNN=98.6%
LR,Adaboost,SVM,DT,	DB	SVM=94.4%
SVM,ANN	Malaria	SVM=89%
DNN	Diabetes	DNN=83.67%

MLP, KNN, CART, SVM, NB.	Breast cancer	MLP=96.70%
NB,LS-SVM, Adabag,Adaboost,	Breast cancer	Adaboost=99.08%
BN,LR,MLP,SMO,DT	Liver cancer	SMO=93.33%
NB,SVM,RF,LR,ANN	Heart disease.	SVM=97.53%
MLP,SVM,KNN,C4.5,RF	Cancer	RF=99.45%
LR,NN,VM	Chronic kidney	VM=97.8%
KNN,SVM,RF, Adaboost	Heart Disease	RF=95.24%
PCA-KNN,PCA-SVM, EM-PCA-Fuzzy Rule-Based	Breast Cancer	EM-PCA-Fuzzy Rule-Based=93.6%
SVM, GEPSVM, TSVM	Alzheimer's	TSVM=92.75%
SVM,L1-Logistic,L2-Logistic,RF,RUSRF	Alzheimer's	SVM=73.33%
RF,SVM,AB,BT,GL	Diabetes	RUSRF=90.60%

Table 4: Techniques are used in Public Health.

Cross Validation Technique

The predictive performance of the models is evaluated using Cross-Validation technique to estimate how each model performs outside the sample to a new dataset also identified as test data. The reason for using cross-validation techniques is to fit it into a training dataset when we fit a model. Cross-validation was applied to achieve the best results in order to measure the numerical performance of a learning operator. This was not achieved to properly isolate and compare the performance of the different methods with respect to the weighting of the propensity score. Through several steps, we measured the quality of the various propensity score matching methods. The classifier's accuracy calculation is the average accuracy of k-folds. Subsampling is done in bootstrap validation with equivalent substitution from the training dataset. Effective use of the 10-fold cross-validation was found to be a good and reasonable compromise between offering accurate performance estimates and being computationally feasible and preventing over fitting [4].

Disease Name	Validation Methods
Parkinson Disease	10 fold
Liver Disease	10 fold
Diabetes Disease	10 fold
Malaria Disease	5 fold
Heart Disease	5 fold
Breast cancer Disease	10 fold
Breast cancer Disease	5 fold
Liver cancer Disease	10 fold
Heart disease	10 fold
Cancer Disease	5 fold
Chronic kidney disease	10 fold
Heart Disease	5 fold
Alzheimer's Disease	10 fold

Table 5: Summary of validation Technique in Public Health.

Model Evaluation Technique

After the estimation, the performance of the predictive models is evaluated in terms of accuracy, accuracy and recall of unseen data using k-fold cross validation technique to test their abilities. Classification performance is evaluated by evaluating the precision, sensitivity and specificity of each system as it is a widely accepted tool of classification performance evaluation and generalization error estimation. It is important to mention that the F1 score can be affected by distorted class ratios when used as a quality indicator. Both AUC and F1 scores are compared using paired t-tests to updated Bonferroni inference thresholds. Here we can summarize different methods of performance evaluation as below

Disease Name	Validation Methods
Parkinson Disease	10 fold
Liver Disease	10 fold
Diabetes Disease	10 fold
Malaria Disease	5 fold
Heart Disease	5 fold
Breast cancer Disease	10 fold
Breast cancer Disease	5 fold
Liver cancer Disease	10 fold
Heart disease	10 fold
Cancer Disease	5 fold
Chronic kidney disease	10 fold
Heart Disease	5 fold
Alzheimer's Disease	10 fold

Table 6: Summary of Performance Evaluation Methods.

Limitations

While the application of machine learning approaches to healthcare problems is unavoidable given the complexity of processing massive amounts of data, the need to standardize standards of interpretable ML in this field is critical Although very broad, these data sets can also be very limited (e.g., system data can only be accessible for a small subset of individuals). Several methods of machine learning effectively address these limitations but are still subject to the usual sources of bias commonly found in experimental studies [5]. The limitation of using SVM is its interpretation, computational costs for larger datasets, and SVM is essentially a binary classifier. Simplified decision tree with four attributes for a multi-class decision problem. A model that is over fitted is more complicated than the data can explain. For genuine disease-related structure, an over fitted model may have too many free parameters and thus risk confusing random noise or other confounding in the training data. This is a pervasive problem in numerical machine learning because it is often possible to set the complexity of the model as high as required to achieve arbitrarily high prediction accuracy. Some of the limitations of traditional medical scoring systems are the presence in the input set of intrinsic linear combinations of variables, and therefore they are not able to model complex nonlinear interactions in medical domains. In this study, this weakness is addressed by using classification models that can implicitly detect complex nonlinear associations between independent and dependent variables as well as the ability to identify any potential correlations between predictor variables.

Conclusion

To inform clinicians and policy makers, systems powered by machine learning will have to deliver results of interest in action through clinical trials or real-world performance observations. Eventually, classification approaches such as clustering and artificial neural networks would require a complete set of experiments. Most of the researcher used the traditional machine learning algorithm to analysis public health data like as SVM, RF, NB, LR, NN, KNN, ANN and DT and 10-fold cross validation provide the better results. But in public health a major challenge is posed by the high volume of healthcare data. This is our big challenges in public health to handle big data. Besides there are a lot of public health researcher facing problems. Most of the problems that have been found in different research paper is classification problems in public health data. Moreover, overfitting and data imbalances are big problems in public health. In our review paper we find some problems and challenges which keep in mind every public health researcher because most of the research paper discussed about these problems and most of the researchers have faced these problems.

References

1. Li, Wei, Chai Yuanbo, Khan Fazlullah, Ullah Jan Syed Rooh, and Sahil Verma, et al. "A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system." *Mob Netw Appl* 26, (2021): 1-19.
2. Panch, Trishan, Szolovits Peter, and Atun Rifat. "Artificial Intelligence, Machine Learning and Health Systems." *J Glob Health*" 8, (2018).
3. Nair, Lekha R, D Shetty Sujala, and D Shetty Siddhanth. "Applying Spark Based Machine Learning Model on Streaming Big Data for Health Status Prediction." *Comput Electr* 65, (2018): 393-399.
4. Kubota, Ken J, A Chen Jason, and Little Max A. "Machine Learning for Large-Scale Wearable Sensor Data in Parkinson's Disease: Concepts, Promises, Pitfalls, and Futures." *Mov Disord* 31, (2016): 1314-1326.
5. Nakajima, Tetsushi, Katsumata Kenji, Kuwabara Hiroshi, and Soya Ryoko, et al. "Urinary Polyamine Biomarker Panels with Machine-Learning Differentiated Colorectal Cancers, Benign Disease, and Healthy Controls." *Int J Mol Sci* 19, (2018): 756.

How to cite this article: Asadullah, MD, Rashid Mamunar, Basu Priyanka and Murad Hossain MD. "Machine Learning in Public Health: A Review of the Problems and Challenges." *J Health Edu Res Dev* 9 (2021) : 27767