Editorial

International Journal of Clinical Rheumatology

Machine Learning in Rheumatology: The Emerging Cutting-Edge Strategy

Abstract

Machine learning (ML) is a computerized analytical technique that is being increasingly used in biomedicine. ML often provides an advantage over ordinary programmed statisticsin the analysis of big and interrelated information. With the increasing availability of large rheumatology biomedical data, numerous studies have employed ML in rheumatology using electronic health records, imaging, or gene expression data. However, the use of ML has its current strengths and weakness in biomedicine. A better understanding of ML and the future application of advanced ML techniques alongside with the increasing availability of medical large data may facilitate the development of meaningful precision medicine for patients with rheumatic diseases (RDs). In this editorial, we describe the principles of ML, discuss examples of ML application in rheumatology, and illustrate the strengths and weakness of ML.

Keywords: Machine learning ● Rheumatology ● Biomedicine

Tamer A Gheita[1*], Nevin Hammam[2]

[1]Rheumatology Department, Kasr Al-Ainy School of Medicine, Cairo University, Cairo, Egypt.

[2]Rheumatology Department, Faculty of Medicine, Assiut University, Assiut, Egypt,

***Author for Correspondence:**

gheitamer@hotmail.com

Received: 25-Mar-2023, Manuscript No. FMIJCR-23-92821; **Editor assigned:** 27-Mar-2023, Pre-QC No. FMIJCR-23-92821 (PQ); **Reviewed:** 10-April-2023, QC No. FMIJCR-23-92821; **Revised:** 13-April-2023, Manuscript No. FMIJCR-23-92821 (R); **Published:** 20-April-2023, **DOI:** 10.37532/1758-4272.2023.18 (4).72-74

Introduction

Machine learning (ML) is a computerized statistics that involves the use of algorithms to discover underlying relationships and/or predict outcomes in high dimensional data that cannot be identified with other statistical techniques. Unlike statistical modeling, ML models are not programmed, but learn from the input data [1]. Big data of patients with rheumatic diseases (RDs) from national registries, electronic health records (EHR), and genetics often makes purely statistical analyses unfeasible. For example, ML is well suited to determine which of many clinical features are important for predicting an outcome. In addition, ML can be used in big data analyses to identify distinct disease phenotypes in complex chronic diseases, such as systemic lupus erythematosus (SLE), in which the patient population is extremely heterogeneous, the conditions evolve over time and multiple factors contribute to disease burden [2]. ML is frequently used in other areas of biomedicine and subtype patient cohorts based on analyzing medical images [3], to predict disease activity [4], drug response [5] and aid in guiding personalized medicine [6]. ML-driven analyses in biomedicine could greatly improve both research and clinical practice in rheumatology.

Types of machine learning

ML algorithms are employed to build predictive models for: classification, regression, or clustering. **Classification** is the prediction of discrete categories of data into a labeled group, known as a 'class'. For example, ML classification models could distinguish patients from controls. **Regression** is the prediction of an outcome, such as model to predict disease flare or response to therapy. **Clustering** is the grouping of similar features into unlabeled groups known as 'clusters', such as clustering of the same gouty patients on the basis of their similar features [2].

ML algorithms for classification, regression and clustering belong to one of four types of learning, including supervised, unsupervised, semi-supervised and reinforcement learning. Supervised and unsupervised algorithms, the most common algorithms, can be differentiated based on the outcome variable

(labeled or unlabeled). For example, a supervised ML model could be applied to predict the disease activity status of patients [4], whereas an unsupervised model could group the same patients based onshared characteristics among patients, such as similar clinical presentation [2].

Machine learning workflow

The general workflow for all ML models involves data preprocessing, model construction, model training, validation and assessment. **Data preprocessing** includes handling missing data (imputation), data transforming (data scaling), and feature selection (selecting the most appropriate variables to use as input). **Model construction**, selection of the most appropriate algorithm, depends on the questions being asked, and on qualities of the input data. Common **classification** algorithms include random forest (RF), gradient boosting machine (GBM), decision trees, support vector machine(SVM), and algorithms for **clustering** include hierarchical clustering and k-means clustering [7]. **Model training** is an iterative process by which the model 'learns' to classify, regress or cluster the outcome variable. Ideally, all ML models are built and improvedusing two independent datasets: the training dataset and the validation dataset [8]. The training dataset is the initial input data on which the model is built. The validation dataset is either a part of the original dataset, or an entirely separate dataset containing data with the same variables. The **validation** step is used to estimate of how accurately the model predicts the outcome. Another validation technique, k-fold cross-validation; the data are split into a number of groups, k, and the model is run times. **Model assessment** includes multiple metrics such as accuracy, sensitivity, and specificity. If the classification problem is binary, these values are often represented using receiver operating characteristic (ROC) curves, and the area under the curve (AUC).

Machine learning use in rheumatic autoimmune diseases

Patient classification using electronic medical records: The existing potential power of ML classification in EHR data has been demonstrated. However, EHRs contain comprehensive information (demographic, clinical features, drug utilization) about individual patients. EHRs are fundamentally noisy, and may result in false results [9]. Thus, a new approach in EHR-based ML begins with language data transformation; natural language processing (NLP) techniques, so that the classifier algorithm can interpret the data [10].

Patient classification using imaging data: ML can be employed to improve the accuracy of imaging-based diagnosis, evaluation and outcome prediction. The accuracy of these models has been assessed based on their agreement with the analysis of an expert rheumatologist and/or radiologist. ML algorithms have been employed aimed at diagnosing or grading RA disease activity based on imaging data [11]. Deep learning has also been used for image analysis in primary Sjögren's syndrome (pSS) to classify the grade of pSS from salivary gland ultrasonography [12].

Risk classification and outcome prediction: ML offers another possible modality for identifying disease risk markers and predict future outcomes such as resulting organ flare or damage. For example, a supervised Extreme Gradient Boosting (XG Boost) model was used to identify risk of vision complications in patients with Behçet's disease (BD). This model incorporated demographic data, laboratory and clinical parameters and was trained on two classes:'controls', which include patients with BD who had no vision complications, and 'cases', which include patients with BD who developed vision complications. The model was trained until the AUC on the labelled data exceeded 0.95.

Predicting treatment response or candidates for treatment: The ability to accurately predict non-responders to specific therapies would be beneficial for patients and clinicians, so that focus could shift to treatments that are more likely to be effective.

Patient clustering to determine disease subtypes: Subtyping of patients is often thought of as major step in achieving precision medicine. However, validating the subtypes, which may have been determined by supervised or clustering models, is difficult. Studies employed clustering using gene expression data are limited. Using simple clinical data for clustering were described and are ideal for forming clinically meaningful patient subsets. For example, patients with juvenile SLE were clustered using k-means clustering [2]. Unsupervised clustering identified four hidden patterns of patients with similar disease manifestations, and further analysis is ongoing.

Limitations of machine learning

Despite its advantages, ML can still be challenging. With the availability of numerous ML algorithms, it is often difficult to determine the most appropriate algorithm without comparison of several models. Some algorithms may be limited because of the size of the dataset, desired outcome, and the quality of available data. In addition, there is potential for ML models to over fit or under

fit the data and, thereby, these models might produce results that cannot be replicated in an unrelated dataset.

Conclusions

The availability of large dataset in rheumatology, and the application of ML to the analysis of these data are beginning to change the landscape of rheumatology research and might lead to a transformation in care by implementation of rapid, effective precision medicine.

References

1. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. 349 (6245), 255–60 (2015).
2. Identifying Distinct Phenotypes of Patients with Juvenile Systemic Lupus Erythematosus: Results from a Cluster. *Analysis by the Egyptian College of Rheumatology (ECR) Study Group* (2018).
3. Unsupervised Cluster Analysis of Clinical and Ultrasound Features Reveals Unique Gout Subtypes: Results from the Egyptian College of Rheumatology (ECR) . *ACR Meeting Abstracts [Internet]* (2023).
4. Kalweit M, Walker UA, Finckh A *et al* . Personalized prediction of disease activity in patients with rheumatoid arthritis using an adaptive deep neural network. *PLoS ONE*. 16(6), e0252289 (2021).
5. Duquesne J, Bouget V, Cournède PH *et al.* Machine learning identifies a profile of inadequate responder to methotrexate in rheumatoid arthritis. *Rheumatology (Oxford)*. keac645 (2022).
6. Peng J, Jury EC, Dönnes P *et al.* Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges. *Front Pharmacol [Internet]*. 12,720694 (2021).
7. Kevin G. Moores, Nila A. Sathe. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data – *ScienceDirect* (2013).
8. Zhao SS, Hong C, Cai T *et al.* Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. *Rheumatology (Oxford) [Internet]*. 59(5), 1059–1065 (2019).
9. Matsuo H, Kamada M, Imamura A *et al.* Machine learning-based prediction of relapse in rheumatoid arthritis patients using data on ultrasound examination and blood test. *Sci Rep [Internet]*. 12(1), 7224 (2022).
10. Vukicevic AM, Radovic M, Zabotti A *et al.* Deep learning segmentation of Primary Sjögren's syndrome affected salivary glands from ultrasonography images. *Comput Biol Med*. 129, 104154 (2021).
11. Hammam N, Bakhiet A, El-Latif EA *et al.* Development of machine learning models for detection of vision threatening Behçet's disease (BD) using Egyptian College of Rheumatology (ECR)-BD cohort. *BMC Med Inform Decis Mak*. 23(1), 37 (2023).
12. Adam G, Rampášek L, Safikhani Z *et al.* Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol*. 4, 19 (2020).