Early detection of multiple sclerosis and the improvement of clinical trials recruitment process with ML methods

Violeta Todorova, Valeri Mladenov
Faculty of Automatics, Technical University of Sofia, Bulgaria; e-mail: valerim@tu-sofia.bg

Abstract: *The purpose of a clinical trial is to evaluate a new treatment or a medical procedure. When medical researchers conduct a trial, they recruit participants with appropriate already existing health problems and medical histories. We describe an expert system based on Machine Learning (ML) algorithms that helps to recognize multiple sclerosis (MS) patients for clinical trials at a very early stage of the disease development. Experiments show that patients can be selected based on common early signs of MS and thus the system can increase the number of selected patients which helps to the clinical trial requirement. It can be beneficial to the patients to recognize the disease in an early development stage.*

Keywords: *Multiple Sclerosis (MS), Clinical Trial Recruitment, MS early detection, Machine Learning (ML)*

1. INTRODUCTION

Recruiting potential participants to research studies such as clinical trials involves three stages: identifying, approaching and obtaining the consent of potential participants to join a study. Patient enrollment is the most time-consuming aspect of the clinical trial process. The leading cause of missed clinical trial deadlines is patient recruitment, taking up to "30 percent of the clinical timeline" [1]. Researchers often rely on healthcare staff, such as doctors and nurses, to identify and approach potential participants. Screening for eligible trial patients can be a tedious and error-prone process. The most time-consuming element knows which patients to review for matching to the eligibility requirements of the clinical trial, which is often a manual process that can take a significant amount of work and time. Each clinical trial possesses specific eligibility criteria also known as inclusion/exclusion criteria. The difficulty arises that even a patient is eligible for a certain clinical trial; the subject might not be at the protocol required stage of the disease. On the other hand, a suitable patient might not choose to participate due to inconvenience in travel, lack of time to participate, out-of-pocket expenses etc. Last but not least, patients might not be even aware of an appropriate for them trial.

We have conducted an experiment to identify appropriate patients for a clinical trial for patients with early stage of multiple sclerosis (MS). The research is focused on identifying patients before the patients have been diagnosed with the disease. The experiment has used AI techniques, like machine learning (ML), to select subjects based on their medical history and the classifying of MS is performed on common early symptoms of the disease.

Multiple sclerosis (MS) is an autoimmune-mediated disorder that affects the central nervous system (CNS). According to the National Multiple Sclerosis Society, MS affects "400,000 people in the United States" and it "appears more frequently in women, in people aged 20 to 50 years, and in people living farther from the equator" [2]. As a chronic autoimmune disorder, MS results in demyelination and neurodegeneration and it often

leads to "severe physical or cognitive incapacitation as well as neurological problems in young adults" [3]. Choosing an effective treatment is crucial to preventing disability. However, response to treatment varies greatly between patients. Because of this, accurate and timely detection of individual response to treatment is an essential requisite of efficient personalized multiple sclerosis therapy [4]. Multiple Sclerosis (MS) progresses at an unpredictable rate, but predictions on the disease course in each patient would be extremely useful to tailor therapy to the individual needs [5]. The natural course of the disease is extremely variable, ranging from extremely mild to very aggressive forms. Given the clinical heterogeneity of multiple sclerosis, reliable prognostic predictors would be of great importance. Several prognostic factors of disability have been described, and some studies have proposed risk scores calculated from demographic and clinical variables collected at disease onset [6, 7]. However, the prediction of the course of MS on the basis of clinical and other supportive data is challenging, and no validated prediction model for the clinical course is currently available.

There have been a number of studies exploring different machine learning (ML) approaches to predict the course and progression of MS as such need arises from the clinical need to improve MS patients' situation [4,5,11]. As the natural course of MS is extremely variable, ranging from extremely mild to very aggressive forms, a main focus of some research has been the development of a personalized prediction model for each MS patient, applying Machine Learning and Big Data techniques [5]. Such studies have shown that large and well-maintained clinical databases can profitably be used to predict the course of MS in individual patients, inciting physicians to devote some effort to archive their data in a format compatible with computer analysis [5]. Exploration have shown that the development of joint physician-patient visualization and decision-making tools may be further enabled using predictive algorithm and that machine learning techniques may be powerful tools for the personalization of MS therapeutic approaches [12].

Early detection of MS is important because it gives us the opportunity to seek treatment and plan for the future. Because disability from MS accumulates over time, "the diagnosis of MS needs to be accurate and occur as early as possible" [8]. The early detection of MS enables clinicians to initiate early therapy, which has been to shown "to delay the onset of disability, and can possibly slow the accumulation of disability and prevent or delay exacerbations" [9]. This research focuses on the use of AI in early detection of MS and the recognition of the disease based on the appearance of early sign of the disease, their tracking and confirmed diagnosis.

The paper is organized as follows. In the next section we discuss the data collection and description of the data. Then in Chapter three we present two type of approaches supervised learning – support vector machines (SVM) and non-supervised – k-means clustering for early MS detection and the conclusion remarks are given in the final chapter.

2. DATA COLLECTION, DATA DESCRIPTION AND DATA CLEANING

The more common symptoms of MS include "sensory disturbances (numbness, tingling, itching, burning), walking difficulties (due to fatigue, weakness, spasticity, loss of balance and tremor), vision problems (diplopia, blurred, and pain on eye movement),

intestinal and urinary system dysfunction (constipation and bladder dysfunction), cognitive and emotional impairment (inability to learn and depression), dizziness" [10].

We have employed randomly generated patients' medical history data that has been used for the training ML algorithms. We applied two types of ML algorithms, namely controlled and uncontrolled training. For controlled training, a set of training data is used to train the algorithm for recognizing the early symptoms of MS, tracking their development and developing a predictive model that can result in determining whether the patient may be at risk of developing of the Council of Ministers. For unsupervised learning, we try to group data so that patients at risk and patients who are not at risk fall into different clusters.

The data used in the research has been provided by Memorial Medical Center. The file contains anonymized clinical information about patients, compliant with the General Data Protection Regulation (G.D.P.R.), along with longitudinal data from their electronic medical records (EMRs). The raw data consists of all the available data up to that moment, coming from clinical records, covering a period of time slightly longer than 10 years, from 2007 until June 2017.

The clinical records of 342 patients were used out of which 183 attending the MS services of the hospital, and 159 non-MS hospital patients. The raw data includes medical history, containing the reported terms (that are the concomitant disease) for the patients and patient's demographics data: patient's age, gender, race, and country of birth.

The extracted data are often incomplete, contain unnecessary or ambiguous information, and suffer disruptions due to noise or pose other difficulties that affect the performance of the predictive models. Thus, it is necessary to pre-process and validate them to avoid future issues. The process of extracting variables out of the patients' data could be long and tedious. However, we used statistical analysis software to derive the needed variables from the raw data and to prepare the data into a format best suitable for our purposes. Missing values of patient ID have been handled by removing the corresponding clinical record. The data has been subset to contain the five most common Reported Terms for the Medical History for the MS patients and similarly for the non-MS patients. Afterwards the data has been transposed to one record per patient and each Reported Term for the Medical History has been set to constitute a separate column. The presence of a Reported Term for the Medical History has been flagged with a '1' and its absence has been marked with '-1'.

The analysis data used for ML classifiers in this paper (SVM and k-means clustering) consist of: the four most common Reported Terms for the Medical History for the MS patient population from the raw data, which are: 'WEAKNESS (LOCALIZED)', 'GENERALIZED FATIGUE', 'HYPERREFLEXIA', 'BLURRED VISION IN ONE OR BOTH EYES', and the five most common Reported Terms for the Medical History for the Non-MS patient population from the raw data, which are: 'HEADACHES', 'MENSTRUAL CRAMPS', 'APPENDICITIS', 'APPENDECTOMY', 'ACNE'. We used these terms as features, i.e. the feature vector for each patient consists of 9 features.

3. AI METHODS FOR EARLY MS DETECION

In this research, we are interested in answering the use AI techniques, like natural language processing (NLP) and machine learning (ML), to automatically analyze clinical

trial eligibility databases and electronic health records (EHRs), to find matches between ongoing clinical trials and appropriate patients, and to have these matches recommended to patients and investigators for a specific medical diagnosis? Using the collected dataset, we apply two main known technique – support vector machines and k-means clustering for study the data for early detection of multiple sclerosis and thus for improvement of clinical trials recruitment process.

3.1. Support Vector Machines (SVM)

SVM is a supervised learning technique that analyze data used for classification and regression analysis. For two classes classification a label with +1 belongs and -1 does not belong to the corresponding class is used. For our case the classes for patients that are used are +1 – with MS symptoms and -1 – without MS symptoms.

The dataset is divided to two subsets for training – 300 patients and for testing 42 patients. The results are obtained by using ML toolbox of Matlab. The results from training are depicted in Figure 1.

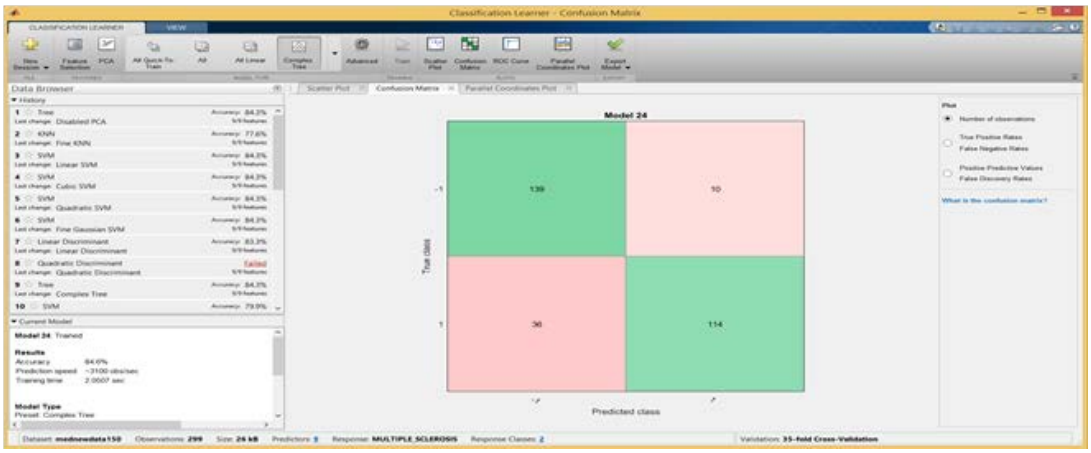

Fig. 1: Learning results for MS detection with different algorithms.

Some other models for supervised learning are used, but the SVM gives the best accuracy of 84.3%. In Fig.2 are given successfully and unsuccessfully learned points. It can be observed that 10 patients (points) are unsuccessfully learned for class -1, whereas the unsuccessfully learned points for class +1 are 35. The percentages of successfully and unsuccessfully learned points for both classes are given in Fig.2.

Based on obtained classifier the all 42 patients used for testing are classified correctly. Furthermore, we did a test removing some features (Reported Terms for the Medical History) and applying the SVM learning. It turns out that removing 'APPENDECTOMY', 'ACNE', 'APPENDICITIS' and 'MENSTRUAL CRAMPS' and based on the other 5 features the accuracy of learning become 84.9%. In the light of the above we can conclude that the reported terms considered should not affect seriously the MS.

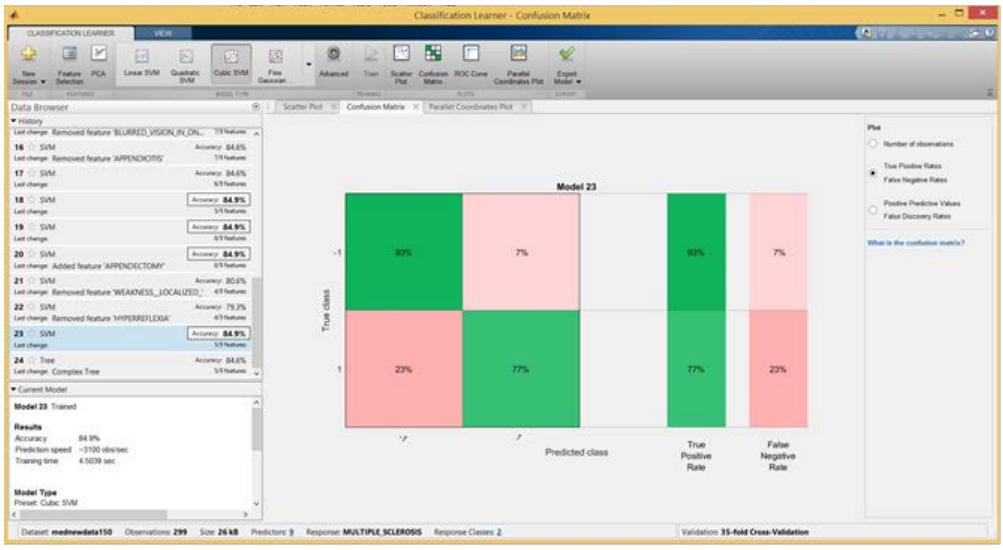

Fig. 2: The percentages of successfully and unsuccessfully learned points for both classes.

3.2. k-means clustering

In non-supervised learning a hidden structure in data where we don't know the right answer upfront is discovered. In the clustering analysis a natural grouping in data is obtained, such that items in the same cluster are more similar to each other than those from different clusters. In k-means clustering each cluster is represented by a "prototype" data point. The clustering in 2, 3 and 4 clusters is shown in Fig 3. According to the mean value of the clusters the most suitable number of clusters is 2. It fits to the fact that the real number of classes that the patients belong to are 2. The first one corresponds to the patients with symptoms for MS and the second one for the patients without MS symptoms.

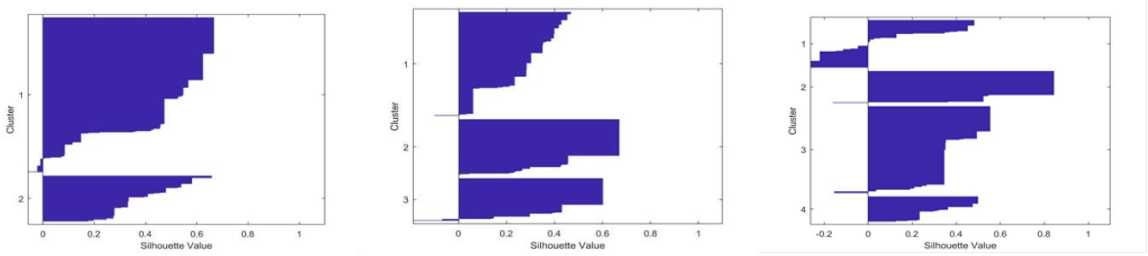

Fig. 3: Clustering in 2, 3 and 4 clusters.

4. CONCLUSIONS

The reported results for application of ML methods to early detection of multiple sclerosis not only can help researches increase the number of selected patients for candidates for a clinical trial but also can be beneficial for patients to be diagnosed with the disease in an early development stage and therefore to have an early start of treatment

since early recognition and accurate diagnosis of MS are crucial to delay disease progression as much as possible and improve patients' outcomes. The results illustrate that the use of ML on patients' medical history data could speed-up the process of clinical trials' patient recruitment and to help find patients who qualify for a specific study.

5. REFERENCES

[1] Bachenheimer, J., Brescia, B., (2007) Reinventing Patient Recruitment: Revolutionary Ideas for Clinical Trial Success.Gower Publishing.

[2] National Multiple Sclerosis Society. Who gets Multiple Sclerosis? http://www.nationalmssociety.org/about-multiple-sclerosis/what-we-know-about-ms/who-gets-ms/index.aspx.

[3] Compston A., Coles A., (2008) Multiple sclerosis. Lancet 372, 1502–1517.

[4] Pruenza, C., (2019) Model for Prediction of Progression in Multiple Sclerosis. Ijimai10.9781.

[5] Seccia, R., (2020) Considering patient clinical history impacts performance of machine learning models in predicting course of multiple sclerosis. Pone Journal 0230219.

[6] Bergamaschi R, Montomoli C, Mallucci(2015) G BREMSO: a simple score to predict early the natural course of multiple sclerosis. Eur J Neurol, 22, 981–989.

[7] Rotstein D, Montalban X, (2019) Reaching an evidence-based prognosis for personalized treatment of multiple sclerosis. Nat Rev Neurol.; 15, 287–300.

[8] Waubant ,E., Improving Outcomes in Multiple Sclerosis Through Early Diagnosis and Effective Management. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583755/.

[9] Kappos L. (2007) Effect of early versus delayed interferon beta-1b treatment on disability after a first clinical event suggestive of multiple sclerosis: a 3-year follow-up analysis of the BENEFIT study. Lancet: 370, 389–397.

[10] Ghasemi, N., Razavi, S., Nikzad, E., Multiple Sclerosis: Pathogenesis, Symptoms, Diagnoses and Cell-Based Therapy, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5241505/.

[11] [11] Zhao, Y, Healy, B. C.,Rotstein, D., (2019) Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data. Proceedings of Machine Learning Research 102:483–492.

[12] Zhao, Y, Healy, B. C.,Rotstein, D., (2017) Exploration of machine learning techniques in predicting multiple sclerosis disease course. Pone Journal 0174866.

ACKNOWLEDGMENTS

The presented results were obtained under a project funded by the research grant at TU-Sofia under project № 202ПД0029-8, "Application of artificial intelligence in recruiting patients for clinical trials" for 2020.