OPEN ACCESS J PREV MED HYG 2021; 62: E192-E199

RESEARCH ARTICLE

Prediction of Multiple sclerosis disease using machine learning classifiers: a comparative study

SONIA DARVISHI[1], OMID HAMIDI[2], JALAL POOROLAJAL[3]

[1] Social Determinants of Health Research Center, Research Institute for Health Development, Kurdistan University of Medical Sciences, Sanandaj, Iran; [2] Department of Science, Hamedan University of Technology, Hamedan, Iran; [3] Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran - Modeling of Noncommunicable Disease Research Center, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

Keywords

Classification • Machine learning • Multiple sclerosis

Summary

Introduction. *Hamedan Province is one of Iran's high-risk regions for Multiple Sclerosis (MS). Early diagnosis of MS based on an accurate system can control the disease. The aim of this study was to compare the performance of four machine learning techniques with traditional methods for predicting MS patients.*
Methods. *The study used information regarding 200 patients through a case-control study conducted in Hamadan, Western Iran, from 2013 to 2015. The performance of six classifiers was used to compare their performance in terms of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR-) and total accuracy.*
Results. *Random Forest (RF) model illustrated better performance among other models in both scenarios. It had greater specificity (0.67), PPV (0.68) and total accuracy (0.68). The most influential diagnostic factors for MS were age, birth season and gender.*
Conclusions. *Our findings showed that despite all the six methods performed almost similarly, the RF model performed slightly better in terms of different criteria in prediction accuracy. Accordingly, this approach is an effective classifier for predicting MS in the early stage and control the disease.*

Introduction

Multiple sclerosis (MS) is a chronic autoimmune inflammatory disease related to the central nervous system (the brain and spinal cord) without clear etiology. Focal lymphocytic infiltration in MS leads to the destruction of myelin and axons [1]. The onset of the disease happens in young adults with the most susceptibility is related to people who are in their 20s and 30s [2, 3]. Approximately 2.5 million people are affected by MS worldwide annually [4]. The prevalence of MS varies geographically between 5.3 and 74.28 per 100,000 in Iran [2]. It has been shown by epidemiological studies that the trend of MS, especially in women, is increasing [2, 5, 6]. Hamadan Province, located in western Iran, is among the most high-risk regions in Iran with the prevalence of 62.5/100,000 [2].

The burden of MS disease for the public health systems and its prevalence during have been increased the past years [7]. Therefore, identifying the most important factors related to the MS is of great importance. Many epidemiologic studies showed that MS has a multifactorial etiology that corresponds to several environmental factors for people who have complex genetic risk profiles [8]. These factors include both genetic and non-genetic exposure to dietary patterns [9], infectious agents [9], familial clustering [10], season of the birth [11], age infection during childhood [11, 12], smoking [13], environment exposures [14], and psychological stress [15].

Early detection of the disease can play a critical role in improving MS survival by increasing the proportion of patients diagnosed at early stages [16]. To do this, traditional classification techniques including logistic regression have been widely used in different medical problems to detect cases and controls. While there can obtain simple interpretations from these models, they usually cannot account for complex relationship between variables. So, the need to use newly developed models with the least prediction error is evident and a precise and reliable system is required to early diagnosis of the patients. Most of modern medical diagnosing tools are constructed based on classification and are adapted by many researchers to improve the precision.

Recently, machine learning techniques have become very popular and have been widely used in several research area including medicine especially in classification problems [17, 18]. These methods learn through experience to improve their performance and can help physicians to better diagnose new patients by increasing sensitivity and in decision-making [19]. Although the main objective of these models is to identify effective variables and their relationships, these models can be used to predict and estimate the effects [20, 21].

Various machine learning methods have been introduced in different studies [22-24]. Examples of them include Naive Bayes (NB), Decision Trees, Random Forest (RF),

https://doi.org/10.15167/2421-4248/jpmh2021.62.1.1651

Nearest Neighbor, AdaBoost, Support Vector Machine (SVM), RBF Network, and Multilayer Perceptron machine learning techniques to predict different outcomes [25-27].

Although different studies have shown that the performance of data mining techniques is better than that of the traditional techniques in terms of higher accuracy and lower error rates, this excellence does not happen in all data sets [28] and there are inconsistencies among various studies. So, investigation and comparison of the performance of different methods in different data sets is of great importance.

The present study aimed to conduct a comprehensive comparison of four machine learning techniques of NB, Least Square Support Vector Machine (LSSVM), SVM and RF and two traditional methods (Logistic Regression (LR) and Linear Discriminant Analysis [29]) in prediction of MS to distinguish people with MS from healthy people in Iran.

Methods

Data source

This study has been approved by the Research Council of the University of Medical Sciences of Hamadan (ID: 9204181211). The data was collected through a case-control study in Hamadan Province, the west of Iran, from September 2013 to March 2014. Participants were voluntarily entered into the study. Due to the lack of intervention, merely verbal informed consent was obtained from the participants. Based on Asadollahi et al. [30], in the patients with MS 80% of the participant was female and in the control group, this percent was 60%. According to this finding, the sample size for each group was 100, the total sample size was 200, at 95% significance level and 80% statistical power. Moreover, 100 definite patients with MS invited to the study as a case group compared to 100 infectious diseases patients as control group who had not a history of neurological disorder. In order to make the study groups similar, individuals from case and control were entered at the same time and in the same hospital. Cases and controls were selected from patients who referred to Farshchian Hospital's neurology clinic and infectious diseases clinic, respectively. The Farshchian Hospital, where the study was conducted, is a referral center to which patients referred from different cities of the province. To make similar the study base of both case and control groups, we decided to select the control group from the Infectious Diseases Ward that was next to the Neurology Ward. Furthermore, the clinical and laboratory information of the control group was available and accessible from their medical records. Regardless of age, gender, and disease onset's date cases were selected. In this study, the individual case was defined as an MS patient who was diagnosed with a neurologist and a brain MRI or a total spinal MRI. The patients with the following criteria were entered to the study: 1) diagnosed during the past 10 years; 2) inhabitant of Hamedan Province; 3) undertreatment and had a complete medical recorder in Farshchian Hospital. Satisfaction and accessibility of patients to study entry was required. The individual control was defined as an infectious disease patient without a neurological disorder seeking medical care. Patients of infectious diseases who have come from other jurisdictions have been disqualified. A standardized questionnaire, embracing of 40 items, was designed for the data collection on socio-demographic characteristics and environmental factors. It included data on gender, age at diagnosis, occupation, marital status, educational level, weight, height, history of smoking, exclusive breastfeeding, history of measles, family history of MS, birth season, history of immune system disease, blood group, and RH variable. The Body Mass Index (BMI), which is the ratio of body weight in kg to height in square meters, was classified into three categories of individuals with BMI underweight (BMI < 18.5), average individuals (BMI = 18.5-24.9), and overweight or obese individuals (BMI ≥ 25). Moreover, to assess the participants' personality type the Friedman-Rosenman standard questionnaire was used. There were 25 two-choice (yes/no) questions with a total score of 25 in the questionnaire. Patients' scores were classified to ≥ 13 and < 13 as type A and type B personality, respectively [31, 32]. The personality questionnaires reliability, used by Cronbach's alpha coefficient, was 0.77. Face-to-face interviews were carried out to collect data.

Data mining algorithms

Naive Bayes (NB)

This classification method is based on the theorem of Bayes, which is straightforward, simple and quick [33, 34]. Once the test and train datasets have been allocated, the prior probability of belonging to each class can be determined using the train set using the conditional probability of independent variables X_i, given the class label C of the output variable. The probability of C is computed the

$$\underset{\sim}{X}_1,\ldots,\underset{\sim}{X}_n \left(P(C = c \mid \underset{\sim}{X}_1 = \underset{\sim}{x}_1,\ldots,\underset{\sim}{X}_n = \underset{\sim}{x}_n) \right.$$

using a class label product probabilities and the conditional probability of independent variables given the class label in theory and based on the Bayes theorem.

$$P(C = c \mid \underset{\sim}{X}_1 = \underset{\sim}{x}_1,\ldots,\underset{\sim}{X}_n = \underset{\sim}{x}_n) = P(C = c) \times \prod_{x_j} P(X_j = x_j \mid C = c)$$

Based on the above formula, the class with the highest posterior probability is given a new event [25].

Support vector machine (SVM)

SVM is a mapping function that uses a classification or regression model that is well known as a flexible method. To perform the classification method, a nonlinear kernel function is implemented to transform independent variables into high dimensional space, in which cases can be differentiated very well. The Radial Basis Function (RBF)

kernel makes a trade-off between the misclassification of the training sample against the simplicity of the decision surface (cost parameter). The outcome variable class is best differentiated by using the maximum-margin hyperplanes in the data. A minimal generalizing error is achieved when the distance between the hyperplanes is accomplished by comparing two parallel hyperplanes on either side of the separating hyperplane [35].

Least Square Support Vector Machine (LSSVM)

The LS-SVM is a modified model with the least squares of the loss function and the equality constrain of the SVM model, in which rather than the quadratic programming problem, the dual solution could be found by solving a linear system. The LS-SVM function, also, maps the data into a high dimensional space, in case of SVM. The primal formulation of the LS-SVM classification model is minimized

$$f(\alpha, c, \varepsilon) = \frac{1}{2}\|\alpha\|^2 + \frac{1}{2}B(\|\varepsilon\|^2)$$

with the equality constraint as:

$$y_i(\alpha^T \varphi(x_i) + c) = 1 - \varepsilon_i, (\varepsilon_i \geq 0)$$ [36].

Random Forest (RF)

The RF method was introduced by Leo Breiman [37] where the regression trees and classification are assembled. In this method, the trees are generated by using a replacement sampling of the main dataset. Using the independent variables that evaluate the outcome and the random subset of the predictors, the nodes are built. The most effective predictors can be found using mean decrease Gini and mean decrease accuracy [37].

Logistic Regression (LR)

This method assumes that the binary outcome is distributed binomially. The model can be written as:

$$\log\left(\frac{\pi}{1-\pi}\right) = \sum_{i=1}^{k} \beta_i X_i$$

In this model, X's are the covariates and β_i is the regression coefficients denoting the effect size's measure [38].

Linear Discriminant Analysis (LDA) [29]

LDA is similar to LR and refers to a linear combination of predictors that can achieve clear interpretations of the dependent variable. LDA addresses the problem with the predictor's conditional probability given the output class. This method maximizes the dispersion between the different class cases and minimizes it between the same class cases [39].

Evaluation criteria and cross validation

To compare the discriminative powers of the classification methods, several criteria of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR-) and total accuracy were calculated using the following formulas:

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \quad \text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{PPV} = \frac{TP}{TP+FP}, \quad \text{NPV} = \frac{TN}{TN+FN}$$

$$\text{LR+} = \frac{\text{Sensitivity}}{\text{1-Specificity}}, \quad \text{LR-} = \frac{\text{1-Sensitivity}}{\text{Specificity}}$$

$$\text{Total Accuracy} = \frac{TP+TN}{TP+TF+TN+FN}$$

Where FP indicates people with MS that were incorrectly identified as healthy, TP stands for patients with MS that were correctly diagnosed as MS, TN stands for healthy controls that were correctly identified as healthy people and FN stands for patients with MS who incorrectly identified as healthy.

The most important variables are chosen to demonstrate how each variable contributes to the uniformity of the nodes and leaves in the resulting RF by its greatest mean Gini decrease [37, 40]. The Gini coefficient for the child nodes is measured and compared to that of the original node each time a particular variable is used for splitting a node. Furthermore, Partial Dependence Plot demonstrates the nature of the dependence of the approximate estimation of function on each explanatory variable. The research has been conducted using RStudio software v 3.6.2.

Results

Data description

The data set included 200 patient records in which 100 definite patients with MS invited to the study as case group compared to 100 infectious disease patients as controls group who had not a history of neurological disorder.

Table I displays the demographic and clinical characteristics of the participants. Females accounted for 80% of cases and 48% of controls ($P < 0.001$). The control group's mean (SD) age was higher than the case group's; 41.2 (14.8) years *vs* 36.1 (11.5) years respectively. Most of the participants were married and had no academic degree. In controls group, the smoking statues ratio was significantly higher than in cases (27 *vs* 6%; $P < 0.001$). Nevertheless, in cases, the number of widows and divorcees was lower than in controls, but there was no statistically significant difference ($P = 0.074$). Breastfeeding in cases was higher in comparison to the controls ($P <0.01$). Moreover, in the cases, patients with a history of measles were lower than in controls ($P <0.05$).

Tab. I. demographic and clinical characteristics of the case and control groups.

Variable	Cases (%)	Controls (%)	P-value
Gender			
Male	20 (20)	52 (52)	0.000
Female	80 (80)	48 (48)	
Marital status			
Single	25 (25)	19 (19)	0.3
Married	75 (75)	81 (81)	
Educational level			
Non-academic	63 (63)	72 (72)	0.1
Academic	37 (37)	28 (28)	
Positive family history			
No	90 (90)	94 (94)	0.2
Yes	10 (10)	6 (6)	
Smoking status			
Non-smoker	94 (94)	73 (73)	0.000
Smoker	6 (6)	27 (27)	
Exclusive breast feeding			
Non-breast feeding	22 (22)	6 (6)	0.001
Breast feeding	78 (78)	94 (94)	
History of measles			
No	64 (64)	50 (50)	0.04
Yes	36 (36)	50 (50)	
Season of birth			
Spring	23 (23)	33 (33)	0.08
Summer	27 (27)	31 (31)	
Autumn	29 (29)	15 (15)	
Winter	21 (21)	21 (21)	
Blood group			
AB	5 (5)	14 (14)	0.3
A	21 (21)	22 (22)	
B	20 (20)	24 (24)	
O	25 (25)	40 (40)	
Blood Rh			
Negative	13 (13)	24 (24)	0.2
Positive	60 (60)	73 (73)	
BMI			
Underweight	2 (2)	4 (4)	0.6
Normal weight	32 (32)	33 (33)	
Overweight & obesity	32 (32)	43 (43)	
Type of personality			
B	30 (30)	40 (40)	0.1
A	70 (70)	60 (60)	

Performance of the models

In order to avoid overfitting, we divided the data into two sets of scenarios including training (70%) and testing set (30%) and training (50%) and testing (50%). We also repeated this process 100 times and reported the evaluation criteria as average over 100 repetitions. Table II provides a comparison of the sensitivity, specificity, PPV, NPV, total accuracy, LR+ and LR- for the classification methods training and test sets. According to the results, in both scenarios, all methods performed quite similarly in terms of LR+ and LR-. Higher accuracy was achieved by the RF in both scenarios and the SVM method in comparison to others.

As the performances of all methods in the classification of MS patients and controls were similar, we calculated the variable importance to rank the role of the variables in predicting MS. According to the results shown in Figure 1, the points represent the mean decrease Gini value, indicative of the importance of each variable in the RF plot, age was the first top rank variable in predicting MS. Also, season and sex were the second and third top rank variables in terms of the Mean decrease in the Gini index. Here, we used a threshold of 10 for Gini index, then we chosethree variables as the most important variable.
Moreover, the partial dependence plot (PDP) of the classes was computed and visualized the relationship

Tab. II. Mean and standard deviation of sensitivity, specificity, PPV, NPV, total accuracy, positive LR and negative LR for various models.

Scenario	Models	Sensitivity		Specificity		PPV		NPV		TA		LR+		LR-	
		Mean	Std. dv	Mean	Std. dv	Mean	Std. dv	Mean	Std. dv	Mean	Std. dv	Mean	Std. dv	Mean	Std. dv
70, 30	NB	0.79	0.10	0.55	0.13	0.64	0.08	0.74	0.10	0.67	0.05	1.92	0.54	0.35	0.14
	LSSVM	0.61	0.09	0.67	0.09	0.65	0.08	0.64	0.07	0.64	0.05	2.06	0.86	0.51	0.01
	RF	0.71	0.08	0.67	0.08	0.68	0.07	0.70	0.08	0.68	0.04	2.06	0.61	0.51	0.11
	SVM	0.72	0.89	0.64	0.1	0.67	0.08	0.70	0.09	0.68	0.05	2.06	0.65	0.51	0.13
	LR	0.67	0.08	0.65	0.1	0.66	0.08	0.66	0.09	0.66	0.06	2.06	0.64	0.51	0.15
	LDA	0.68	0.08	0.64	0.10	0.66	0.08	0.66	0.09	0.55	0.06	2.06	0.66	0.51	0.16
50, 50	NB	0.77	0.15	0.56	0.15	0.64	0.07	0.74	0.10	0.66	0.05	1.90	0.45	0.37	0.17
	LSSVM	0.62	0.10	0.63	0.10	0.63	0.07	0.63	0.06	0.63	0.04	1.90	0.46	0.51	0.13
	RF	0.71	0.09	0.57	0.09	0.68	0.07	0.70	0.07	0.68	0.04	1.90	0.43	0.51	0.11
	SVM	0.69	0.10	0.63	0.1	0.65	0.06	0.68	0.08	0.66	0.04	1.90	0.42	0.51	0.13
	LR	0.67	0.09	0.63	0.08	0.65	0.06	0.66	0.07	0.66	0.04	1.90	0.65	0.51	0.11
	LDA	0.68	0.09	0.63	0.09	0.64	0.06	0.66	0.07	0.65	0.04	1.90	0.50	0.51	0.12

Fig. 1. Variable importance in predicting MS disease using RF model.

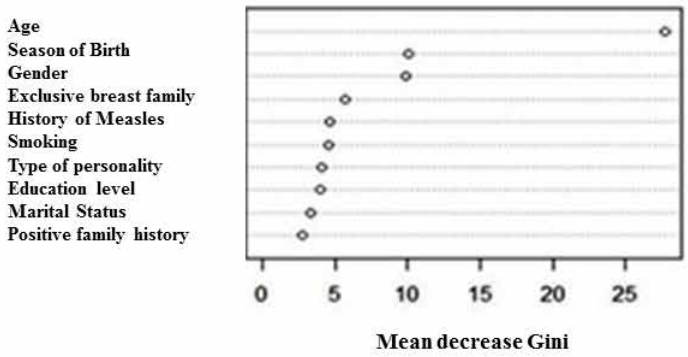
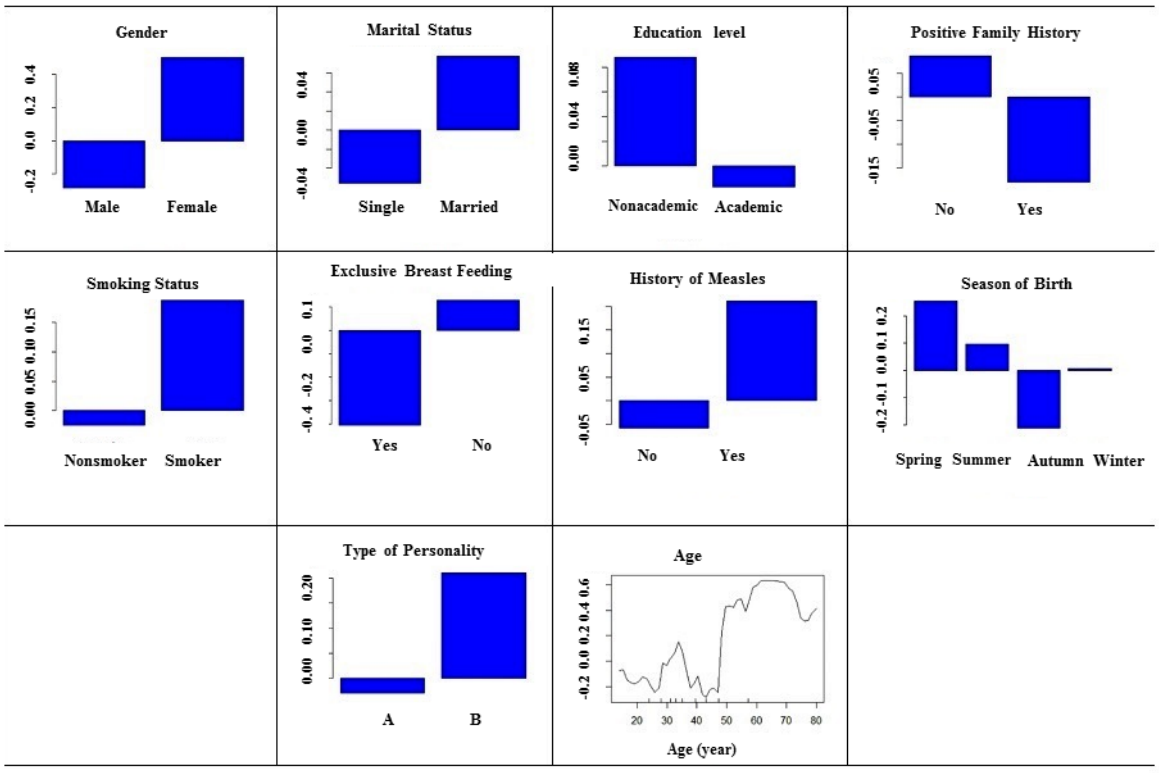


Fig. 2. Partial plots for variables in predicting MS using RF.



between prediction of MS on different features for the RF. Figure 2 shows that there is an MS prediction for female, married, non-academic education, history of measles, birth in spring, history of smoking and b personality type.

Discussion

The present study was aimed at a comprehensive comparison of six machine learning techniques of NB, LSSVM, SVM, RF and two traditional methods (LR and LDA) for the prediction of MS to distinguish people with MS from healthy people in Iran.

For all six methods, the performance criteria were very similar among classifiers, however, they derived from different algorithm approaches. Based on the total accuracy, it was shown that in both scenarios: 1) 70% training and 30% testing; and 2) 50% training, 50% testing), all classification methods performed almost the same for the classification of MS cases and controls (ranged: 0.54 to 0.68). Only, one of the six classifiers tested showed a total accuracy value lower than 0.6 (LDA with total accuracy of 0.55). In other words, in predicting the classes for both case and control groups, all the classification methods provided similar accuracy. However, the total accuracy of the RF model was slightly more than others in both scenarios (0.68).

In the 70, 30 scenarios, the sensitivity varied from at least 0.61 in LSSVM to at most 0.79 in the NB model. This indicator is also accurate in 50, 50 scenarios (0.77 in NB model). In the case of specificity, however, the RF model performed better than other models (0.67), the NB model was poor (0.55). This quality also remains true for PPV. In other words, RF is the best model based on the NPV and PPV criteria.

The maximum sensitivity and NPV value belonged to NB. However, the RF model outperformed other models on the basis of the other reliability indices and it is more effective than NB, LSSVM, SVM, LR and LDA. Moreover, RF and NB showed similar accuracies. Since, they were the most common algorithms used in practice [29, 30, 41, 42], RF model was used for additional analysis.

Our finding indicated age as the highest risk factor associated with MS prediction. This result is consistent with the findings [1, 7, 43]. MS is more likely to occur in the 20-40 age group [1, 7, 43]. Our analysis indicates patients in their late 20 to mid-30 were at a high risk of MS. The PDP showed that the predicted MS probability is low until 50 increases after. The result of previous studies was inconsistent with this finding [44] .

According to the finding, season on birth was the second important variable in predicting MS patients, consistent with previous findings [11]. The PDP presented that the MS risk in patients who born in spring and summer was more common. Cruz et al in the United Kingdom also founded that spring-born patients are at greater risk than autumn-born patients [45]. Walleczek et al study also found a significant rise in MS births in April and a decline in November [46]. On the other hand, our analysis opposed the results of some previous literature that reported autumn-born patients had a higher risk of MS than spring [1, 47]. In 2019, a systematic survey and multivariate meta-analysis was conducted to address this conflict and revealed that in the northern hemisphere, the impact of the birth season was related to latitude, annual dry bulb temperature and sunshine period. For populations in latitudes > 52° this impact was restricted to the sunshine period [48] .

The third factor that influences the prediction was gender. According to a PDP, the probability of having MS is more likely to be diagnosed in females than in males. Our finding was performed the similar result of preceding research [3, 49, 50]. This can be due to the disparity between women and men in the immune state, nervous system, and lifestyle in both sexes [3]. The propensity to have fewer children and have them later in life than their grandmothers is one of the big changes in the life of the contemporary woman. Due to temporary immunosuppressant during pregnancy [51], pregnancy can have a protective impact against MS in women, and a higher age may have a share of the increased incidence of MS in women when giving birth to the first child or fewer pregnancies [51].

There were several limitations to our study. Firstly, in order to establish the models, we did not focus on quantitative MRI features. Further work plans to incorporate additional biomarker data. Second, there was some limitation in the number of samples and the matching of age and sex in both cases and control groups.

Conclusions

The aim of this research was to evaluate the performance of four machine learning and two classical techniques in predicting MS patients. Our findings suggest that in this study, RF was the best model for predicting MS in terms of multiple criteria between two group patients.

Acknowledgements

Funding sources: this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

We would like to thank the Vice-chancellor of Research and Technology of Hamadan University of Medical Sciences for financial support of this study. The authors declare that there is no conflict of interests.

Conflicts of interest statement

The authors declare no conflict of interest.

Authors' contributions

The research topic was constructed by SD and OH, the idea was investigated, the statistical analysis

was performed and the manuscript was drafted. JP provided the data and participated in the preparation of interpretations and manuscripts. The final manuscript was read and approved by all authors.

References

[1] Poorolajal J, Mazdeh M, Saatchi M, Ghane ET, Biderafsh A, Lotfi B, Feryadres M, Pajohi K. Multiple sclerosis associated risk factors: a case-control study. Iran J Public Health 2015;44:1498. https://doi.org/10.1007/s10072-003-0147-6

[2] Etemadifar M, Sajjadi S, Nasr Z, Firoozeei TS, Abtahi SH, Akbari M, Fereidan-Esfahani M. Epidemiology of multiple sclerosis in Iran: a systematic review. Eur Neurol 2013;70:356-63. https://doi.org/10.1159/000355140

[3] Harbo HF, Gold R, Tintoré M. Sex and gender issues in multiple sclerosis. Ther Adv Neurol Diso 2013;6:237-48. https://doi.org/10.1177/1756285613488434

[4] Di Cara M, Lo Buono V, Corallo F, Cannistraci C, Rifici C, Sessa E, D'Aleo G, Bramanti P, Marino S. Body image in multiple sclerosis patients: a descriptive review. Neurol Sci 2019;40:923-8. https://doi.org/10.1007/s10072-019-3722-1

[5] Moghtaderi A, Rakhshanizadeh F, Shahraki-Ibrahimi S. Incidence and prevalence of multiple sclerosis in southeastern Iran. Clin Neurol Neurosurg 2013;115:304-8. https://doi.org/10.1016/j.clineuro.2012.05.032

[6] Rezaali S, Khalilnezhad A, Naser Moghadasi A, Chaibakhsh S, Sahraian MA. Epidemiology of multiple sclerosis in Qom: demographic study in Iran. Iran J Neurol 2013;12:136-43. https://www.ncbi.nlm.nih.gov/pubmed/24250923

[7] Etemadifar M, Abtahi SH. Multiple sclerosis in Isfahan, Iran: Past, Present and Future. Int J Prev Med 2012;3:301-2. https://doi.org/10.1159/000094235

[8] Sellner J, Kraus J, Awad A, Milo R, Hemmer B, Stuve O. The increasing incidence and prevalence of female multiple sclerosis - a critical analysis of potential environmental factors. Autoimmun Rev 2011;10:495-502. https://doi.org/10.1016/j.autrev.2011.02.006

[9] Ramsaransing GS, Mellema SA, De Keyser J. Dietary patterns in clinical subtypes of multiple sclerosis: an exploratory study. Nutr J 2009;8:36. https://doi.org/10.1186/1475-2891-8-36

[10] Guaschino C, Esposito F, Liberatore G, Colombo B, Annovazzi P, D'Amico E, Cavalla P, Capello E, Capra R, Galimberti D, Tedeschi G, Grimaldi L, Progresso Group, Progemus Group, Leone M, D'Alfonso S, Martinelli V, Comi G, Martinelli-Boneschi F. Familial clustering in Italian progressive-onset and bout-onset multiple sclerosis. Neurol Sci 2014;35:789-91. https://doi.org/10.1007/s10072-014-1650-7

[11] Vazirinejad R, Sotoudeh-Maram E, Soltanzadeh AA, Taghavi M-M. The effect of childhood viral infections on the incidence of multiple sclerosis. Zahedan J Res Med Sci 2013;15:24-7.

[12] Mikaeloff Y, Caridade G, Suissa S, Tardieu M, Group KS. Clinically observed chickenpox and the risk of childhood-onset multiple sclerosis. Am J Epidemiol 2009;169:1260-6. https://doi.org/10.1093/aje/kwp039

[13] Handel AE, Williamson AJ, Disanto G, Dobson R, Giovannoni G, Ramagopalan SV. Smoking and multiple sclerosis: an updated meta-analysis. PLoS One 2011;6:e16149. https://doi.org/10.1371/journal.pone.0016149

[14] Orton SM, Wald L, Confavreux C, Vukusic S, Krohn JP, Ramagopalan SV, Herrera BM, SadovnickAD, Ebers GC. Association of UV radiation with multiple sclerosis prevalence and sex ratio in France. Neurology 2011;76:425-31. https://doi.org/10.1212/WNL.0b013e31820a0a9f

[15] Pakenham KI. Making sense of illness or disability: the nature of sense making in multiple sclerosis (MS). J Health Psychol 2008;13:93-105. https://doi.org/10.1177/1359105307084315

[16] World Health Organization. Breast cancer: breast cancer and early diagnosis. Available from: http://wwwwhoint/cancer/prevention/diagnosis-screening/breast-cancer/en/.2018

[17] Zhao Y, Healy BC, Rotstein D, Guttmann CR, Bakshi R, Weiner HL, Brodley CE, Chitnis T. Exploration of machine learning techniques in predicting multiple sclerosis disease course. PLoS One 2017;12:e0174866. https://doi.org/10.1371/journal.pone.0174866

[18] Boucekine M, Loundou A, Baumstarck K, Minaya-Flores P, Pelletier J, Ghattas N, Auquier P. Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. BMC Med Res Methodol 2013;13:20. https://doi.org/10.1186/1471-2288-13-20

[19] Ramana BV, Babu MSP, Venkateswarlu N. A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems 2011;3:101-14. https://doi.org/10.5121/ijdms.2011.3207

[20] Hashemian AH, Beiranvand B, Rezaei M, Bardideh A, Zand-Karimi E. Comparison of artificial neural networks and cox regression models in prediction of kidney transplant survival. Int J Adv Biol Biomed Res 2013;1:1204-12.

[21] A novel memetic feature selection algorithm. Information and Knowledge Technology (IKT), 2013 5th Conference on. Montazeri M, Montazeri M, Naji HR, Faraahi A (eds.). IEEE 2013.

[22] Das R, Sengur A. Evaluation of ensemble methods for diagnosing of valvular heart disease. Expert Syst Appl 2010;37:5110-5. https://doi.org/10.1016/j.eswa.2009.12.085

[23] Alkim E, Gurbuz E, Kilic E. A fast and adaptive automated disease diagnosis method with an innovative neural network model. Neural Netw 2012;33:88-96. https://doi.org/10.1016/j.neunet.2012.04.010

[24] Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 2014;41:1476-82. https://doi.org/10.1016/j.eswa.2013.08.044

[25] Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technol Health Care 2016;24:31-42. https://doi.org/10.3233/THC-151071

[26] Chao C-M, Yu Y-W, Cheng B-W, Kuo Y-L. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst 2014;38:106. https://doi.org/10.1007/s10916-014-0106-1

[27] Ferrer L, Rondeau V, Dignam J, Pickles T, Jacqmin-Gadda H, Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. Stat Med 2016;35:3933-48. https://doi.org/10.1002/sim.6972

[28] Finch H, Schneider MK. Classification accuracy of neural networks vs discriminant analysis, logistic regression, and classification and regression trees. Methodology 2007;3:47-57.

[29] El Khoury Y, Collongues N, De Sèze J, Gulsari V, Patte-Mensah C, Marcou G, Varnek A, Mensah-Nyagan AG, Hellwig P. Serum-based differentiation between multiple sclerosis and amyotrophic lateral sclerosis by Random Forest classification of FT-IR spectra. Analyst 2019;144:4647-52. https://doi.org/10.1039/c9an00754g

[30] Asadollahi S, Fakhri M, Heidari K, Zandieh A, Vafaee R, Mansouri B. Cigarette smoking and associated risk of multiple sclerosis in the Iranian population. J Clin Neurosci 2013;20:1747-50. https://doi.org/10.1016/j.jocn.2013.01.018

[31] Friedman M, Rosenman RH. Type A behavior and your heart. Fawcett 1974.

[32] Shaygannejad V, Dehnavi SR, Ashtari F, Karimi S, Dehghani L, Meamar R, Tolou-Ghamari Z. Study of type a and B behavior patterns in patients with multiple sclerosis in an Iranian population. Int J Prev Med 2013;4(Suppl 2):S279-83. https://www.ncbi.nlm.nih.gov/pubmed/23776738

[33] Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. Age. 2006;58:10-110.

[34] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record 2002;31:76-7.

[35] Auria L, Moro RA. Support vector machines (SVM) as a technique for solvency analysis (August 1, 2008). DIW Berlin Discussion Paper No. 811. Available at SSRN: https://ssrn.com/abstract=1424949

[36] Tripathy RK, Zamora-Mendez A, de la O S, José A, Paternina MRA, Arrieta JG, Naik GR. Detection of life threatening ventricular arrhythmia using digital taylor fourier transform. Front Physiol 2018;9:722. https://doi.org/10.3389/fphys.2018.00722

[37] Breiman L. Random forests. Mach Learn 2001;45:5-32.

[38] Agresti A. Categorical data analysis. John Wiley & Sons 2003.

[39] Izenman A. Linear discriminant analysis. Modern multivariate statistical techniques. New York: Springer 2013.

[40] Yitzhaki S, Schechtman E. The Gini methodology: a primer on a statistical methodology. Springer Science & Business Media 2012.

[41] Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform 2019;125:55-61. https://doi.org/10.1016/j.ijmedinf.2019.02.002

[42] Risk prediction of type II diabetes based on random forest model. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). Xu W, Zhang J, Zhang Q, Wei X (eds.). IEEE 2017.

[43] Tullman MJ. Overview of the epidemiology, diagnosis, and disease progression associated with multiple sclerosis. Am J Manag Care 2013;19(Suppl 2):S15-20. https://www.ncbi.nlm.nih.gov/pubmed/23544716

[44] Sanai SA, Saini V, Benedict RH, Zivadinov R, Teter BE, Ramanathan M, Weinstock-Guttman B. Aging and multiple sclerosis. Mult Scler 2016;22:717-25. https://doi.org/10.1177/1352458516634871

[45] Cruz PMR, Matthews L, Boggild M, Cavey A, Constantinescu CS, Evangelou N, Giovannoni G, Gray O, Hawkins S, Nicholas R. Time-and region-specific season of birth effects in multiple sclerosis in the United Kingdom. JAMA Neurol 2016;73:954-60. https://doi.org/10.1001/jamaneurol.2016.1463

[46] Walleczek NK, Frommlet F, Bsteh G, Eggers C, Rauschka H, Koppi S, Assar H, Ehling R, Birkl C, Salhofer-Polanyi S, Baumgartner A, Blechinger S, Buchinger D, Sellner J, Kraus J, Moser H, Mayr M, Guger M, Rathmaier S, Raber B, Liendl H, Hiller MS, Parigger S, Morgenstern G, Kempf I, Spiss HK, Meister B, Heine M, Cisar A, Bachler H, Khalil M, Fuchs S, Enzinger C, Fazekas F, Leutmezer F, Berger T, Kristoferitsch W, Aboulenein-Djamshidian F. Month-of-birth-effect in multiple sclerosis in Austria. Mult Scler 2019;25:1870-7. https://doi.org/10.1177/1352458518810924

[47] Staples J, Ponsonby AL, Lim L. Low maternal exposure to ultraviolet radiation in pregnancy, month of birth, and risk of multiple sclerosis in offspring: longitudinal analysis. BMJ 2010;340:c1640. https://doi.org/10.1136/bmj.c1640

[48] Pantavou KG, Bagos PG. Season of birth and multiple sclerosis: a systematic review and multivariate meta-analysis. J Neurol 2019:1-8. https://doi.org/10.1007/s00415-019-09346-5

[49] Kearns PKA, Paton M, O'Neill M, Waters C, Colville S, McDonald J, Young IJB, Pugh D, O'Riordan J, Weller B, MacDougall N, Clemens T, Dibben C, Wilson JF, Castro MC, Ascherio A, Chandran S, Connick P. Regional variation in the incidence rate and sex ratio of multiple sclerosis in Scotland 2010-2017: findings from the Scottish Multiple Sclerosis Register. J Neurol 2019;266:2376-86. https://doi.org/10.1007/s00415-019-09413-x

[50] Leray E, Moreau T, Fromont A, Edan G. Epidemiology of multiple sclerosis. Rev Neurol 2016;172:3-13. https://doi.org/10.1016/j.neurol.2015.10.006

[51] McCombe PA, Greer JM. Female reproductive issues in multiple sclerosis. Mult Scler 2013;19:392-402. https://doi.org/10.1177/1352458512452331

Received on June 25, 2020. Accepted on February 16, 2021.

Correspondence: Omid Hamidi, Department of Science, Hamedan University of Technology, Hamedan, 65155 Iran - Tel.: +98 81 38411533 - E-mail: omid_hamidi@hut.ac.ir

How to cite this article: Darvishi S, Hamidi O, Poorolajal J. Prediction of Multiple sclerosis disease using machine learning classifiers: a comparative study. J Prev Med Hyg 2021;62:E192-E199. https://doi.org/10.15167/2421-4248/jpmh2021.62.1.1651