Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2023;12(4):1315-9

The role of advanced machine learning approach in predicting multiple sclerosis development and progression

Mehmet Ediz Sarihan[1], Zeynep Kucukakcali[2]

[1]*İnönü University Faculty of Medicine, Department of Emergency Medicine, Malatya, Türkiye*
[2]*İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye*

Received 19 September 2023; Accepted 12 October 2023
Available online 24.10.2023 with doi: 10.5455/medscience.2023.09.196

Abstract

This study aims to utilize a fine-tuned gradient boosting trees algorithm to predict the onset and progression of Multiple Sclerosis (MS) based on a comprehensive set of demographic and clinical variables. The goal was to enhance early diagnosis and enable individualized treatment approaches using artificial intelligence. The research utilized Dataset, a publicly accessible dataset derived from a prospective cohort study of individuals of Mexican mestizo descent diagnosed with Clinically Isolated Syndrome (CIS). The study spanned from 2006 to 2010, systematically collecting and analyzing data on various individual traits to explore correlations with MS development. The gradient boosting trees algorithm was employed to construct predictive models, harnessing patient-specific variables, including demographic factors and clinical data. The classifier exhibited outstanding performance, with a mean accuracy of 99.63% and minimal standard deviation. The confusion matrix indicated one false positive and no false negatives. Key metrics such as precision, recall, and AUC all approached 1, demonstrating the classifier's ability to distinguish between the two classes with high confidence. Comparative analysis with similar studies in the literature revealed superior performance, highlighting the classifier's accuracy and effectiveness in predicting MS. The application of the gradient boosting trees algorithm to predict MS based on demographic and clinical variables offers a promising avenue for early diagnosis and tailored treatment. This research demonstrates the potential of AI to transform healthcare, particularly in the context of MS. The predictive models developed have the capacity to enhance early detection, improve patient quality of life, and pave the way for further AI-based solutions in healthcare.

Keywords: Multiple sclerosis, gradient boosting trees, artificial intelligence, healthcare innovation

Introduction

A common disabling condition that affects young adults without a history of trauma is multiple sclerosis (MS). The number of people diagnosed with MS is increasing in both developed and developing countries, but the exact reason for this trend is unclear [1,2]. MS is a complex condition, with many genes slightly increasing the risk of developing the disease, as well as several known environmental factors, such as vitamin D or ultraviolet B light (UVB) exposure, infection with the Epstein-Barr virus (EBV), obesity, and smoking. MS was traditionally considered as an autoimmune disease caused by T-cells that attack specific organs, but this view has been challenged by the success of treatments that target B-cells, which contradicts the conventional T-cell autoimmune theory [3,4]. MS is usually seen as a two-phase condition, with initial inflammation causing relapsing-remitting disease, followed by later neurodegeneration leading to non-relapsing progression, called secondary and primary progressive MS [5,6].

MS is a complicated disease that arises from the interplay of various genetic variants and external influences that have an impact on its vulnerability. These genetic and environmental factors combine to initiate the development of this disorder. It is important to note that comprehensive analyses from observational studies emphasize the importance of certain environmental risk factors that are linked to MS, such as obesity, lack of vitamin D, infection by the Epstein-Barr virus, and smoking habits. These environmental and lifestyle factors can be modified to prevent or delay the onset of MS, so it is crucial to establish definite causal relationships between these factors and the occurrence of MS [7,8]. To prevent MS and its negative impacts on people and healthcare systems, we need to understand the complex connections between different factors that influence the disease.

CITATION
Sarihan ME, Kucukakcali Z. The role of advanced machine learning approach in predicting multiple sclerosis development and progression. Med Science. 2023;12(4):1315-9.

Corresponding Author: Zeynep Kucukakcali, İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
Email: zeynep.tunc@inonu.edu.tr

This requires a collaborative effort from different fields of science, such as genetics, immunology, epidemiology, and molecular biology. By doing so, we can gain a holistic understanding of the causes of MS and develop specific interventions that can lower the risk of getting the disease. This is a crucial goal as we face the challenges of scientific advancement and public health improvement, and we need to unravel the complicated web of genetic and environmental factors that make people more prone to MS [9,10].

Gradient Boosted Trees (GBT) is a machine learning technique that combines multiple weak learners, usually decision trees, to create a strong learner that can make accurate predictions for regression and classification tasks. GBT works by iteratively adding new trees that fit the residual error of the previous trees, thus reducing the overall error of the model. GBT can optimize any differentiable loss function and can handle various types of data, such as numerical, categorical, or text. GBT is often considered one of the most powerful and versatile machine learning methods [11,12].

The goal of this research is to apply a fine-tuned gradient boosting trees algorithm to forecast MS based on demographic and clinical features, possibly enhancing early diagnosis and individualized treatment approaches from an artificial intelligence viewpoint.

Material and Methods

The study's primary objective was to leverage a publicly accessible dataset to make predictions regarding the status of Multiple Sclerosis (MS) by closely examining a diverse set of individual traits. An open-access data set was used in the study and the current study is a retrospective case-control study. The STROBE (Strengthening the reporting of observational studies in epidemiology) standard was used to assess the risk of bias and the overall quality of this study. This research endeavor not only contributes significantly to the advancement of early prediction methodologies but also facilitates a more profound understanding of the intricate interactions between individual characteristics and the development of MS. The dataset in question originated from a prospective cohort study conducted within a population of individuals of Mexican mestizo descent. These individuals had recently received a diagnosis of Clinically Isolated Syndrome (CIS), a neurological condition, and had proactively sought medical attention at the prestigious National Institute of Neurology and Neurosurgery (NINN) situated in the vibrant city of Mexico City, Mexico. This comprehensive research initiative spanned a substantial period, ranging from 2006 to 2010. During this extensive timeframe, data was systematically collected and subjected to meticulous analysis to investigate various dimensions related to this specific patient cohort and their lived experiences with CIS. In the pursuit of its research objectives, this study undertook a comprehensive examination of various individual traits. Its aim was to uncover potential correlations or patterns that could provide insights into the likelihood of MS development among individuals diagnosed with CIS. By harnessing the knowledge derived from this publicly available dataset, the study sought to offer valuable insights that could substantially enhance the early identification and management of MS.[13] This, in turn, holds the promise of significantly improving the overall quality of life for individuals affected by this challenging neurological condition. In essence, this research initiative not only utilized accessible data but also engaged in a rigorous exploration of individual attributes within a specific patient population. Its ultimate goal was to shed light on the predictive factors associated with MS, thereby contributing to more effective strategies for early intervention and ultimately offering hope and improved outcomes for those impacted by this condition [13]. Information on predictor and target attributes is provided in Table 1.

Table 1. Information and explanations on predictor and target attributes

Predictor	Definition	Role	Full description of the estimator
Age	Age of the patient (in years)	Predictor	Patient age
Schooling	Time the patient spent in school (in years)	Predictor	Patient schooling
Gender	1=male, 2=female	Predictor	Patient gender
Breastfeeding	1=yes, 2=no, 3=unknown	Predictor	Patient breastfeeding status
Varicella	1=positive, 2=negative, 3=unknown	Predictor	Patient varicella status
Initial_Symptoms	1=visual, 2=sensory, 3=motor, 4=other, 5= visual and sensory, 6=visual and motor, 7=visual and others, 8=sensory and motor, 9=sensory and other, 10=motor and other, 11=Visual, sensory and motor, 12=visual, sensory and other, 13=Visual, motor and other, 14=Sensory, motor and other, 15=visual, sensory, motor and other	Predictor	Patient initial symptoms
Mono_or_Polysymptomatic	1=monosymptomatic, 2=polysymptomatic, 3=unknown	Predictor	Patient mono or polysymptomatic status
Oligoclonal_Bands	0=negative, 1=positive, 2=unknown	Predictor	Oligoclonal bands test result
LLSSEP	0=negative, 1=positive	Predictor	Lower limb somatosensory evoked potentials test result
ULSSEP	0=negative, 1=positive	Predictor	Upper limb somatosensory evoked potentials test result
VEP	0=negative, 1=positive	Predictor	Visual evoked potentials test result
BAEP	0=negative, 1=positive	Predictor	Brainstem auditory evoked potentials test result
Periventricular_MRI	0=negative, 1=positive	Predictor	Periventricular magnetic resonance imaging test result
Cortical_MRI	0=negative, 1=positive	Predictor	Cortical magnetic resonance imaging test result
Infratentorial_MRI	0=negative, 1=positive	Predictor	Infratentorial magnetic resonance imaging test result
Spinal_Cord_MRI	0=negative, 1=positive	Predictor	Spinal cord magnetic resonance imaging test result
Group	1=CDMS, 2=non-CDMS	Target	Patient group

DOI: 10.5455/medscience.2023.09.196

GBT starts with an initial model, which can be a constant or a simple tree, and then adds new trees that are trained on the negative gradient of the loss function with respect to the current model's predictions. This means that each new tree tries to correct the mistakes of the previous model, by moving in the direction that minimizes the loss. The new trees are added with a shrinkage factor, which controls the learning rate and prevents overfitting. The shrinkage factor can be tuned as a hyper-parameter of the algorithm. GBT is effective because it can adapt to different types of data and problems, by choosing an appropriate loss function and tree structure. For example, for regression problems, GBT can use a squared error or a Huber loss function, which are robust to outliers. For classification problems, GBT can use a logistic or a multinomial loss function, which can handle binary or multiclass labels. For text data, GBT can use a word embedding or a bag-of-words representation, which can capture semantic and syntactic features. GBT can also use different types of decision trees, such as CART, C4.5, or random forest, which can vary in their splitting criteria, pruning methods, and randomness [14].

In the current research, the GBT model was constructed on the related public dataset using a five-fold cross-validation approach. The Optimize Parameters (Grid) Operator was employed for the learning rate hyper-parameter of the model in predicting MS based on demographic and clinical features. RapidMiner software was utilized for all calculations and modeling tasks. Performance evaluation was performed through Accuracy, Kappa, Area under the ROC curve, Precision, Recall, Lift, F-measure, Sensitivity, Specificity, Youden Index, Positive Predictive Value and Negative Predictive Value.

Results

The performance metrics of the classifier are reported in Table 2 using the classification/confusion matrix, which provides various measures of accuracy, precision, recall, and other statistics.

Table 2. The performance metrics of the gradient boosted trees classifier

Performance Metrics	Value +/- Standard Deviation
Accuracy	99.63% +/- 0.83%
Kappa	0.993 +/- 0.017
The area under the ROC Curve	0.996 +/- 0.009
Precision	99.33% +/- 1.49%
Recall	100.00% +/- 0.00%
Lift	183.24% +/- 2.20%
F-measure	99.66% +/- 0.76%
Sensitivity	100.00% +/- 0.00%
Specificity	99.20% +/- 1.79%
Youden Index	0.992 +/- 0.018
Positive Predictive Value	99.33% +/- 1.49%
Negative Predictive Value	100.00% +/- 0.00%

The results are based on a 10-fold cross-validation with a stratified sampling of the data. The performance metrics are computed for each fold and then averaged over the folds, with the standard deviation as a measure of variability. The performance metrics are also computed for the micro average, which is the overall performance of the classifier on all the data. The results show that the classifier achieves a high level of accuracy, with a mean value of 99.63% and a low standard deviation of 0.83%. The confusion matrix shows that the classifier correctly classifies 124 instances of class 1 and 148 instances of class 2, with only one misclassification of class 2 as class 1. The kappa statistic, which measures the agreement between the classifier and the true labels, is also very high, with a mean value of 0.993 and a low standard deviation of 0.017.

The performance metrics for class 2 as the positive class are also very high, indicating that the classifier is able to distinguish between the two classes with a high degree of confidence. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve, which measures the trade-off between the true positive rate and the false positive rate, is close to 1, with a mean value of 0.996 and a low standard deviation of 0.009. The precision, which measures the proportion of true positives among all predicted positives, is also close to 1, with a mean value of 0.9933 and a low standard deviation of 0.0149. The recall, which measures the proportion of true positives among all actual positives, is equal to 1, with no variability across the folds. The lift, which measures how much better the classifier is than a random guess, is also very high, with a mean value of 183.24% and a low standard deviation of 2.20%. The f-measure, which is the harmonic mean of precision and recall, is also close to 1, with a mean value of 0.9966 and a low standard deviation of 0.0076. The sensitivity, which is another name for recall, is equal to 1, with no variability across the folds. The specificity, which measures the proportion of true negatives among all actual negatives, is also very high, with a mean value of 0.9920 and a low standard deviation of 0.0179. The Youden index, which measures the difference between sensitivity and (1-specificity), is also close to 1, with a mean value of 0.992 and a low standard deviation of 0.018. The positive predictive value (PPV), which is another name for precision, is also close to 1, with a mean value of 0.9933 and a low standard deviation of 0.0149. The negative predictive value (NPV), which measures the proportion of true negatives among all predicted negatives, is equal to 1, with no variability across the folds.

The study was conducted using an open access dataset. Therefore, a post hoc power analysis was conducted at the end of the study. According to the post hoc power analysis using the schooling variable, a power of 0.992 (1-β) was obtained with an estimated effect size of 0.37 based on the independent sample t-test with sample sizes of 125 and 148 in the groups, taking into account the type I error (α) of 0.05.

Discussion

The primary aim of this research is to employ a highly refined gradient boosting trees algorithm for the purpose of predicting the onset and progression of Multiple Sclerosis (MS) by utilizing an extensive array of demographic and clinical characteristics.

This research represents a significant stride in the field of artificial intelligence (AI) within the domain of healthcare, with the overarching objective of bolstering early detection capabilities and facilitating the development of tailored treatment strategies for MS. Central to this research is the application of the gradient boosting trees algorithm, a robust machine learning technique known for its proficiency in handling complex datasets and discerning intricate relationships within them. This algorithm is uniquely suited to analyze a wide spectrum of patient-specific variables, encompassing demographic elements such as age, gender, ethnicity, and geographic location, alongside an intricate tapestry of clinical data, including medical history, diagnostic results, and treatment responses. Through the systematic integration of these diverse data points, the algorithm endeavors to construct robust predictive models capable of forecasting the likelihood of MS development and its progression. The significance of early diagnosis in MS cannot be overstated, as it underpins timely therapeutic interventions, profoundly influencing patient outcomes and the overall course of the disease. Therefore, this research represents a pivotal step toward revolutionizing clinical practice by introducing a data-driven, AI-centric framework. This framework not only enhances the accuracy and timeliness of MS diagnosis but also empowers healthcare providers to design customized treatment approaches, thereby elevating the quality of care and improving the prospects of individuals affected by MS. In essence, this research embodies a promising convergence of AI and healthcare, poised to reshape the landscape of MS diagnosis and treatment through its data-driven and individualized approach.

The performance findings consists of various metrics that measure the accuracy, precision, recall, sensitivity, specificity, and other aspects of the classifier's performance. The results show that the classifier achieved a very high accuracy of 99.63% with a low standard deviation of 0.83%, indicating that the classifier was able to correctly classify almost all the instances in the data set. The confusion matrix also shows that there were only one false positive and zero false negatives, meaning that the classifier did not misclassify any instances of class 1 as class 2 or vice versa. The kappa statistic, which measures the agreement between the classifier and the true labels beyond chance, was also very high at 0.993 with a low standard deviation of 0.017, indicating that the classifier's performance was not due to random guessing. The performance values also includes the area under the curve (AUC) metric, which measures the ability of the classifier to discriminate between the two classes. The AUC value was 0.996 with a low standard deviation of 0.009, indicating that the classifier had a very high discriminative power and could separate the two classes very well. The positive class in this case was class 2, which means that the classifier was more interested in identifying instances of class 2 than class 1. The precision, recall, lift, f-measure, sensitivity, specificity, youden index, positive predictive value, and negative predictive value metrics all measure different aspects of the classifier's performance for the positive class. All these metrics were very high, ranging from 99.20% to 100%, with low standard deviations, indicating that the classifier was very effective in identifying instances of class 2 and did not miss any or make many errors.

There have been many studies on MS based on clinical and image data. With the use and successful results of machine learning methods in the field of health, the effects of machine learning methods in MS studies are also noticeable. In one study, a machine learning model was created to analyze clinical data for the detection of MS disease progression. The aim of the study was to distinguish between benign and progressive structures. The machine learning methods used in the study were ordinary least squares (OLS), regularized least squares (RLS), K-nearest neighbors (KNN), logistic regression (LR) and linear support vector machines [15]. Another study utilized machine learning to create a clinical decision support system to identify MS patients noninvasively. The study was based on image processing and used an ANN model with a 'Tan-sigmoid' transfer function. used. Feature extraction was based on the identification of features that differ significantly between MS patients and healthy subjects. The study achieved 92.35% accuracy with ANN [16]. In another study, it was proposed to classify serum lipid biomarker classification with serum lipid data for MS using Random Forest, one of the machine learning models. The study achieved 100% accuracy. However, when the study was analyzed, it was seen that there was an unbalanced class problem in the data [17]. Compared to similar studies in the literature, the classifier in this study also achieves superior performance in terms of various metrics measuring accuracy, precision, recall, sensitivity, specificity and discriminative power to predict the class labels of the dataset.

The implications of these results are that they provide evidence for the validity and reliability of the classifier as a tool for predicting the class labels of the data set based on its features. The results also suggest that the features that were used to train and test the classifier were relevant and informative for distinguishing between the two classes.

The main limitations of this study are that it used only one type of classifier and one type of data set to evaluate its performance vector. Future studies could use different types of classifiers such as neural networks or decision trees and different types of data sets such as text or image data to compare their performance vectors with those obtained by this study. Future studies could also explore different ways to optimize or improve the performance vector of the classifier such as feature selection or parameter tuning.

Conclusion

This research has demonstrated the feasibility and efficacy of applying a gradient boosting trees algorithm to predict the onset and progression of MS using a comprehensive set of demographic and clinical variables. The algorithm has shown remarkable performance in capturing the complex and multifaceted nature of MS, as well as its heterogeneous manifestations across different

DOI: 10.5455/medscience.2023.09.196

individuals. By harnessing the power of AI and machine learning, this research has contributed to advancing the field of healthcare, particularly in relation to MS diagnosis and management. The proposed predictive models have the potential to facilitate early detection of MS, which is crucial for initiating appropriate and timely treatment. Moreover, the models can enable healthcare providers to tailor treatment strategies according to the specific characteristics and needs of each patient, thereby improving the quality of life and prognosis of those living with MS. This research represents a novel and innovative approach to tackling a challenging and debilitating disease, paving the way for further exploration and refinement of AI-based solutions in healthcare.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
Since this study was conducted using an open access data set, an ethics committee certificate was not obtained.

References

1. Dimitrov LG, Turner B. What's new in multiple sclerosis? Br J Gen Pract. 2014;64:612-3.

2. Nicholas R, Rashid W. Multiple sclerosis. Am Fam Physician. 2013;87:712.

3. Alfredsson L, Olsson T. Lifestyle and environmental factors in multiple sclerosis. Cold Spring Harbor perspectives in medicine. 2019;9:1-12.

4. Waubant E, Lucas R, Mowry E, et al. Environmental and genetic risk factors for MS: an integrated review. ANN CLIN TRANSL NEUR. 2019;6:1905-22.

5. Dobson R, Giovannoni G. Multiple sclerosis–a review. Eur J Neurol. 2019;26:27-40.

6. Cree BAC, Arnold DL, Chataway J, et al. Secondary progressive multiple sclerosis: new insights. Neurology. 2021;97:378-88.

7. Hone L, Jacobs BM, Marshall C, et al. Age-specific effects of childhood body mass index on multiple sclerosis risk. J Neurol. 2022;269:5052-60.

8. Dyment DA, Ebers GC, Sadovnick AD. Genetics of multiple sclerosis. The Lancet Neurology. 2004;3:104-10.

9. Vandebergh M, Degryse N, Dubois B, Goris A. Environmental risk factors in multiple sclerosis: Bridging Mendelian randomization and observational studies. J Neurol. 2022;269:4565-74.

10. Price E, Lucas R, Lane J. Experiences of healthcare for people living with multiple sclerosis and their healthcare professionals. Health expectations : an international journal of public participation in health care and health policy. 2021;24:2047-56.

11. Delgado-Panadero Á, Hernández-Lorca B, García-Ordás MT, Benítez-Andrades JA. Implementing local-explainability in gradient boosting trees: Feature contribution. Information Sciences. 2022;589:199-212.

12. Ebrahimi M, Mohammadi-Dehcheshmeh M, Ebrahimie E, Petrovski KR. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models. Comput Biol Med. 2019;114:103456.

13. Chavarria V, Espinosa-Ramírez G, Sotelo J, et al. Conversion Predictors of Clinically Isolated Syndrome to Multiple Sclerosis in Mexican Patients: A Prospective Study. Arch Med Res. 2023:102843.

14. Louk MHL, Tama BA. Dual-IDS: A bagging-based gradient boosting decision tree model for network anomaly intrusion detection system. Expert Syst Appl. 2023;213:119030.

15. Fiorini S, Verri A, Tacchino A, Ponzio M, Brichetto G, Barla A, editors. A machine learning pipeline for multiple sclerosis course detection from clinical scales and patient reported outcomes. 2015 37th Annual International Conference of the IEEE engineering in medicine and biology society (EMBC); 2015: IEEE.

16. Sarbaz Y, Pourakbari H, Vojudi MH, Ghanbari A. Introducing a decision support system for multiple sclerosis based on postural tremor: a hope for separation of people who might be affected by multiple sclerosis in the future. Biomedical Engineering: Applications, Basis and Communications. 2017;29:1750046.

17. Lötsch J, Schiffmann S, Schmitz K, et al. Machine-learning based lipid mediator serum concentration patterns allow identification of multiple sclerosis patients with high accuracy. Sci Rep. 2018;8:14884.