Advances in Artificial Intelligence and Machine Learning; Research 4 (1) 2027-2051 Received 6-01-2024; Accepted 10-03-2024; Published 17-03-2024

Machine Learning Algorithms for Early Prediction of Multiple Sclerosis Progression: A Comparative Study

Kamel-Dine Haouam k_haouam@yu.edu.sa
Computer Engineering Department,
College of Engineering and Architecture, Al Yamamah University
Riyadh 13541, Saudi Arabia

Mourad Benmalek
Computer Engineering Department,
College of Engineering and Architecture, Al Yamamah University
Riyadh 13541, Saudi Arabia

Corresponding Author: Kamel-Dine Haouam

Abstract

Multiple Sclerosis (MS) is a chronic autoimmune disease characterized by central nervous system (CNS) degeneration, leading to diverse neurological symptoms. Managing MS poses a challenge due to its unpredictable progression. This study focuses on early prediction of MS progression using machine learning (ML) algorithms, comparing the effectiveness of Random Forest, XGBoost, Decision Tree, and Logistic Regression. Clinical, genetic, and environmental factors were analyzed in a cohort of Mexican mestizo patients recently diagnosed with Clinically Isolated Syndrome (CIS). Data preprocessing addressed missing values, and feature selection tailored to the population's characteristics was applied. The dataset was split into training and testing sets, maintaining stratification for CDMS and non-CDMS cases.

Machine learning models were trained with optimized hyperparameters. Performance evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, were employed. The Random Forest model exhibited superior performance (AUC: 0.93, accuracy: 87%), outperforming other models. Variable importance analysis identified top predictors, including Periventricular MRI, Age, Infratentorial MRI, and Oligoclonal Bands. The study's clinical implications highlight ML's potential in enhancing early MS prognosis, aiding timely interventions. The Random Forest model, with its robust performance, emerges as a valuable tool for identifying patients at risk.

While the study contributes to predictive analytics in neurological disorders, limitations include cohort specificity and retrospective data use. Prospective studies and further exploration of data sources are recommended for broader applicability.

This research demonstrates the efficacy of ML in early MS progression prediction, providing clinicians with a promising tool for personalized patient care. The findings contribute to

Citation: Kamel-Dine Haouam and Mourad Benmalek. Machine Learning Algorithms for Early Prediction of Multiple Sclerosis Progression: A Comparative Study. Advances in Artificial Intelligence and Machine Learning. 2024;4(1):116.

advancing predictive healthcare analytics and emphasize the significance of tailored interventions in neurological disorders.

Keywords: Multiple Sclerosis (MS), Clinically Isolated Syndrome (CIS), Machine Learning (ML), Predictive analytics, Random forest, XGBoost, Decision tree, Logistic regression, Early prognosis, Disease progression, Mexican mestizo, Central nervous system (CNS), Variable importance, Feature selection.

1. INTRODUCTION

Multiple Sclerosis (MS) is a chronic autoimmune disease characterized by the degeneration of the central nervous system (CNS), leading to a wide range of neurological symptoms. The most common neurological inflammatory illness that disables young individuals is multiple sclerosis (MS) [1, 2]. In particular, MS is defined by persistent inflammation that damages White Matter (WM) myelin, causing sclerotic lesions to develop, mostly in the Central Nervous System (CNS) [3]. One of the major challenges in managing MS is the unpredictable nature of its progression, with patients experiencing varying degrees of disability over time. According to Zhang et al. [4], a single episode of neurological symptoms suggestive of an inflammatory demyelinating illness of the central nervous system is referred to as a Clinically Isolated Syndrome (CIS). Early prediction of MS progression is crucial for timely intervention and personalized treatment strategies. Recent advancements in machine learning (ML) techniques have opened new avenues for predicting disease outcomes, enabling healthcare providers to make informed decisions and improve patient outcomes. Walton et al. [5], report that the condition affects around 2.8 million people globally, or one in 3,000 people who have MS. Numerous incapacitating symptoms, including ataxia, sensory impairment, cognitive dysfunction, and weariness, have been reported to be caused by it [6]. Relapsing-remitting MS (RRMS), Clinical Isolated Syndrome (CIS), and progressive MS (PrMS) are the three phenotypes of MS [6].

In the context of MS, predicting disease progression involves analyzing an intricate interplay of clinical, genetic, and environmental factors. In this study, we focus on the early prediction of MS progression, specifically comparing the effectiveness of various machine learning algorithms in forecasting the progression of the disease in its early stages. The literature has demonstrated and extensively supported the use of Machine Learning (ML) techniques to effectively handle the analysis of high-dimensional data for model-informed drug development [7, 8], these techniques have also been used to predict patients' MS disease course and conversion to secondary progressive MS [9, 10]. Our research explores the integration of machine learning models with a comprehensive set of independent variables, including Varicella exposure, brainstem auditory evoked potentials (BAEPs), Visual evoked potential (VEP), Oligoclonal bands (OCBs), somatosensory evoked potentials (SSEP), Expanded Disability Status Scale (EDSS), and specific MRI findings (Periventricular, Cortical, Infratentorial, and Spinal Cord)

Traditionally, diagnosing and predicting the progression of MS have heavily relied on clinical observations and a limited set of biomarkers. The advent of machine learning algorithms has revolutionized the field of medical diagnostics by enabling the analysis of large and complex datasets. New technologies like artificial intelligence and machine learning can analyze multidimensional data to find groups with similar properties, according to Eshaghi et al. [11]. Artificial intelligence and machine learning, for example, hold enormous potential for classifying patients based not on shared clinical traits but rather on similar pathobiological mechanisms when applied to apparent anomalies on MRI scans [12]. Artificial intelligence (AI) approaches have demonstrated their significance in industrial machinery, innovative tool design, and digital technology in recent years [13]. According to Dobson and Giovannoni [14], the combination of clinical symptoms and results from magnetic resonance imaging (MRI) forms the current diagnostic strategy. By harnessing the power of ML, we aim to enhance the accuracy and reliability of early MS progression predictions. This research is not only significant for its potential impact on patient care but also for advancing the methodologies employed in predictive healthcare analytics.

Purpose

This research endeavors to address a critical need within the medical community by introducing a pioneering approach to predicting the progression of MS. It holds immense importance for the medical community, offering a novel approach to predicting MS progression. Early identification of MS progression can lead to timely therapeutic interventions, thereby improving the quality of life for patients and potentially delaying the onset of severe disabilities. Furthermore, the study contributes to the growing body of literature on the application of machine learning in healthcare, paving the way for future research in predictive analytics for neurological disorders. Classification is the primary category of machine learning methods. These categorization methods may be used for a wide range of health issues, including illness screening, diagnosis, and prediction. Clinical data analytics include a variety of machine learning techniques, including artificial neural networks (ANN), Random Forests, k-nearest neighbors (KNN), support vector machines (SVM), decision trees (DT), and Naive Bayes (NB).

The subsequent sections and their respective contents are outlined as follows:

Literature Review: This section will provide an in-depth review of existing literature surrounding Multiple Sclerosis (MS) prognosis, predictive analytics, and the application of machine learning techniques in healthcare. It will explore the current challenges in managing MS progression, the traditional diagnostic and predictive approaches, and recent advancements in machine learning methodologies for early disease prognosis.

Methodology: In this section, the research methodology employed in the study will be elucidated. It will detail the machine learning algorithms utilized, the dataset utilized for analysis, feature selection techniques, model evaluation criteria, and the rationale behind the comparative analysis of different machine learning models.

Results & Findings: The findings of the study, including the performance evaluation of various machine learning algorithms in predicting MS progression, will be presented and discussed in this section. It will highlight the effectiveness of different algorithms, identify key predictors of MS progression, and discuss any insights gleaned from the analysis.

Conclusion: The conclusion section will summarize the key findings of the study, discuss their implications for clinical practice and future research directions, and offer recommendations for leveraging machine learning techniques in improving early prediction and management of MS progression. It will also emphasize the significance of this research in advancing the field of predictive healthcare analytics and its potential impact on patient care.

2. LITERATURE REVIEW

In a research with thirty healthy volunteers and forty-two relapsing-remitting MS patients, Damasceno et al. [15], discovered a correlation between early MS inflammatory disease activity and subsequent clinical impairment. For both clinical deterioration and minimal activity of disease activity, the accuracy of the relationship varied from 70.6% to 71.4%. The utilized methodology relied on logistic regression; however, other machine learning approaches that may produce better results were overlooked.

In the study by Tousignant et al. [16], an innovative deep learning framework is introduced for predicting future disability progression in patients with MS using multi-modal brain Magnetic Resonance Images (MRI). Employing parallel convolutional pathways inspired by Inception net, the model achieves significant accuracy on large clinical trial datasets. With an AUC of 0.66 ± 0.055 using baseline MRI and an improved AUC of 0.701 ± 0.027 when incorporating lesion label masks, the model provides valuable predictions for clinicians. The incorporation of uncertainty estimates based on Monte Carlo dropout sample variance enhances the model's practicality by assisting clinicians in identifying scans requiring further examination.

In order to predict the course of the illness, Brichetto et al. [17], used machine learning to patient-reported (PROs) and clinical-assessed outcomes (CAOs) using a dataset of 3398 evaluations from 810 MS patients. Their research predicts the course with an accuracy rating of 82.6%. Raeisi et al. [18], conducted an MS study in which they examined and classified the disease's clinical symptoms using machine learning techniques. The findings indicated that the most influential factors on MS were clinical symptoms related to vision. They opined that many MS patients often ignore the illness's transient symptoms.

Storelli et al. [19], proposed a deep learning algorithm utilizing baseline T2-weighted and T1-weighted brain MRI scans to predict disease worsening in MS patients after a 2-year follow-up. The algorithm demonstrated high accuracy in predicting clinical (83.3%) and cognitive (67.7%) worsening. The highest accuracy (85.7%) was achieved when incorporating both clinical (Expanded Disability Status Scale) and cognitive (Symbol Digit Modalities Test) information. The model outperformed two expert physicians (70% accuracy).

Eshaghi et al. [20], employed unsupervised machine learning on MRI data from 6322 MS patients to identify subtypes based on pathological features. They classified MS into cortex-led, normal-appearing white matter-led, and lesion-led subtypes. The lesion-led subtype exhibited the highest risk of disability progression and relapse rate, with positive treatment responses in selected trials. Validation on a cohort of 3068 patients supported subtype differentiation. The study suggests MRI-based subtypes can predict disability progression and treatment response in MS, aiding patient stratification in interventional trials.

Zhao et al. [21], employed machine learning techniques to predict MS disease course using data from the CLIMB and EPIC datasets. Utilizing three popular algorithms and three ensemble methods, they achieved AUC scores of 0.79 and 0.83 for CLIMB and EPIC datasets, respectively. XGBoost and LightGBM ensemble models outperformed standalone algorithms, with EDSS, Pyramidal Function, and Ambulatory Index identified as key predictors. This study highlights the efficacy of ensemble learning in forecasting MS progression shortly after onset, providing valuable insights for improved disease course prediction.

In their study, Law et al. [22], employed machine learning (ML) algorithms to predict disability progression in secondary progressive MS (SPMS) participants (n=485). Three decision tree (DT)-based models outperformed logistic regression (LR) and support vector machines (SVMs), achieving higher area under the receiver operating characteristic curve (AUC) values (61.8%, 60.7%, and 60.2% vs. 49.5%-51.1%). The findings suggest that non-parametric ML approaches are superior in predicting SPMS disability progression. This has implications for refining clinical trial cohorts, potentially minimizing low-risk individuals' exposure to experimental therapies.

Vázquez-Marrufo et al. [20], conducted a systematic review exploring the application of machine-learning (ML) algorithms in MS. Utilizing the PubMed search engine, they identified 76 relevant articles, excluding non-English or non-Spanish studies and those not specifically related to MS. The review categorized four main applications of ML in MS: 1) classifying MS subtypes; 2) distinguishing MS patients from healthy controls and individuals with other diseases; 3) predicting progression and treatment responses; and 4) other applications. The findings suggest that ML algorithms have significant potential to support health professionals in both clinical and research settings concerning MS.

Andorra et al. [23], utilized machine learning algorithms to analyze a prospective cohort of 322 MS patients and 98 healthy controls, incorporating clinical, imaging, and omics data. Their study successfully predicted outcomes such as disability accumulation, no evidence of disease activity (NEDA), immunotherapy onset, and therapy escalation with intermediate to high accuracy. The algorithms demonstrated comparable performance in an independent cohort of 271 MS patients. The integration of clinical, imaging, and omics data through machine learning aids in identifying MS patients at risk of disability worsening.

Xia et al. [24], developed Disease Severity in MS using electronic health records (EHR) and identified 5495 MS patients using a natural language processing method. This analysis was based on the ”brain parenchymal fraction” (BPF) and the ”MS severity score” (MSSS). The algorithm demonstrated 83% sensitivity and 95% specificity. Wang et al. [25], employed data mining techniques to identify and manage MS. Fast diagnosis and treatment are made easier by applying data mining techniques with high percentages of sensitivity and specificity, the findings show.

Zhao et al. [26], employed machine learning techniques, specifically support vector machines (SVM), to predict the disease course in MS patients. Utilizing data from the CLIMB study, the authors incorporated short-term clinical observations and MRI data over one year, achieving improved predictive accuracy. SVM, with non-uniform misclassification costs favoring sensitivity, demonstrated promising results, outperforming logistic regression. Key predictors included race, family history of MS, brain parenchymal fraction for non-worsening cases, and brain T2 lesion volume for worsening cases. The study suggests the potential of SVM for guiding treatment decisions in MS.

In a recent study by Plati et al. [27], the authors addressed the MS severity estimation problem and predicted disease progression using Machine Learning (ML) techniques. Utilizing data from the ProMiSi project, encompassing demographic details, clinical data, test results, treatment, and comorbidities of 30 patients recorded at three time points, the ML methods achieved notable accuracy, with 94.87% for MS severity estimation and 83.33% for disease progression prediction. This research, presented at the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, contributes valuable insights to the application of ML in managing MS.

Machine learning methods that estimate an individual's risk of MS were created by a team at Siemens Healthineers in Tarrytown, New York, under supervision from Raj Gopalan, MD. The Beth Israel Deaconess Medical Center in Boston provided them with approximately 3,000 sets of electronic medical records, including those of individuals with and without MS. They analyze patient data that was gathered up to three years before to the diagnosis, including age, gender, blood indicators, and metabolic information, using a ”random forest model”. Their random forest model showed great predictive ability and was incredibly accurate. The model primarily relied on blood measures of neutrophils, red blood cells, and other indicators to identify individuals who were considered to be at-risk.

Ramanujam et al. [28] developed a decision tree classifier to accurately assign disease phenotypes in MS based on clinical data from 14,387 patients in Sweden. The classifier, utilizing the most recent expanded disability scale status score and age, achieved 89.3% accuracy, with validation in an independent cohort yielding 82.0% accuracy. This outperformed a previously published algorithm and demonstrated comparable accuracy to neurologists' clinical judgment. The model's potential to standardize phenotype definitions across cohorts suggests its utility in assisting neurologists by providing valuable additional information, emphasizing its clinical relevance in managing MS patients.

In a study conducted at the National Institute of Neurology and Neurosurgery (NINN) in Mexico City, Mexico, researchers gathered data from Mexican mestizo patients recently diagnosed with Clinically Isolated Syndrome (CIS). Preliminary assessments revealed an intriguing gap in the existing research: a lack of diverse machine learning algorithms employed to predict whether a patient's condition would progress to Clinically Definite MS (CDMS) or remain classified as Non-CDMS. Addressing this gap, our study focused on utilizing various machine learning methods to predict the progression of the disease in these patients. Specifically, we aimed to determine whether a patient's condition would be classified as Clinically Definite MS, thereby contributing valuable insights to the field and enhancing our understanding of disease prognosis in this population.

3. METHODOLOGY

3.1 Data Collection and Preprocessing:

The research focused on gathering data from Mexican mestizo patients recently diagnosed with Clinically Isolated Syndrome (CIS) at the National Institute of Neurology and Neurosurgery (NINN) in Mexico City, Mexico. The collected data included demographic information, detailed medical history, neurological examination results, and other relevant clinical variables specific to the Mexican mestizo population. The dataset underwent rigorous preprocessing, addressing missing values and inconsistencies. Feature selection techniques were applied to identify pertinent variables tailored to the unique characteristics of Mexican mestizo patients.

3.1.1 Data splitting

The preprocessed dataset was divided into training and testing sets, maintaining the stratification of CDMS and Non-CDMS cases within the Mexican mestizo cohort. This approach ensured that the machine learning models were trained and evaluated on data representative of the population under study. The training set enabled the models to learn the intricacies of CIS progression specific to Mexican mestizo patients.

3.2 Machine Learning Algorithms

Four machine learning algorithms - Random Forest, XGBoost, Decision Tree, and Logistic Regression - were chosen for this study due to their suitability for classification tasks. These algorithms were implemented with initial default hyperparameters.

3.2.1 Random forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Each tree is constructed using a subset of the data and a random selection of features, reducing the risk of overfitting and increasing accuracy. In this study, Random Forest can effectively capture complex relationships within the unique genetic and clinical features of Mexican mestizo patients. Random Forest is well-suited for this study due to its ability to handle high-dimensional data, non-linear relationships, and interactions among features. It can efficiently handle both numerical and categorical variables, making it versatile for capturing diverse patient characteristics in the CIS population.

3.2.2 XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithm. It sequentially builds multiple weak learners (typically decision trees) and combines them to create a strong predictive model. XGBoost is renowned for its speed, performance, and ability to handle complex datasets with high dimensionality, making it suitable for large-scale studies like this. XGBoost's ability to handle missing data and its flexibility in objective functions make it valuable in this research. By optimizing its hyperparameters, XGBoost can effectively capture subtle patterns within the data, enabling accurate predictions specific to the Mexican mestizo population with CIS.

3.2.3 Decision tree

Decision Tree is a simple yet powerful algorithm that recursively splits the dataset into subsets based on the most significant features, creating a tree-like structure. It's easy to interpret and visualize, providing valuable insights into the decision-making process. In this study, Decision Trees can highlight key factors contributing to the progression of CIS to CDMS among Mexican mestizo patients. Decision Trees are beneficial when there are interactions and non-linear relationships within the data. They can handle both numerical and categorical variables, making them suitable for capturing various aspects of the Mexican mestizo patient population. Decision Trees can also identify important features, aiding in understanding the underlying mechanisms of CIS progression.

3.2.4 Logistic regression

Logistic Regression is a statistical method used for binary classification tasks. It models the probability of a binary outcome (in this case, CIS progression to CDMS) based on predictor variables. Logistic Regression assumes a linear relationship between the predictors and the log-odds of the binary outcome. It's interpretable and provides insight into the influence of each predictor variable on the outcome. Logistic Regression is valuable in this study as it offers a clear understanding of the impact of individual features on CIS progression. It can handle both numerical and categorical predictors, making it applicable to the diverse dataset of Mexican mestizo patients. While it might not capture complex interactions as effectively as ensemble methods, its interpretability and simplicity are advantageous for gaining insights into the early stages of MS progression in this specific population.

3.3 Hyperparameter Tuning

Hyperparameter tuning was performed using techniques such as grid search with cross-validation, considering the specific genetic and environmental factors relevant to the Mexican mestizo population. Special attention was given to identifying hyperparameters that accounted for the population-specific nuances, ensuring the models were optimized for accurate predictions in this context.

3.4 Model Training and Evaluation:

The selected machine learning algorithms were trained on the training dataset, emphasizing the unique features and characteristics of the population. During training, the models learned patterns specific to this population, enhancing their ability to make accurate predictions regarding the progression of CIS to Clinically Definite MS (CDMS).

The models' performance was evaluated on the test data set by using a range of metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Evaluation results were thoroughly analyzed to understand the models' performance in predicting CIS progression among Mexican mestizo patients. The performance of each model on the test data was examined and the final model selected is the model with best performance metrics.

4. RESULTS & FINDINGS

When an individual seeks evaluation n following a singular episode of Central Nervous System (CNS) inflammation, various decisions regarding long-term follow-up must be considered. Among these is the crucial determination of whether to initiate therapy. Consequently, there is a growing imperative to pinpoint predictive factors that can anticipate the progression from a Clinically Isolated Syndrome (CIS) to MS. In the scope of this endeavor, we aim to concentrate on the early prognosis of MS progression, with a particular emphasis on evaluating the efficacy of diverse machine learning algorithms in predicting the advancement of the disease during its initial phases. To predict the target variable Group (1=CDMS, 2=non-CDMS), we employed four machine learning models: random forest, Xgboost, decision tree, and logistic regression. The machine learning pipeline for this classification is presented as FIGURE 1.

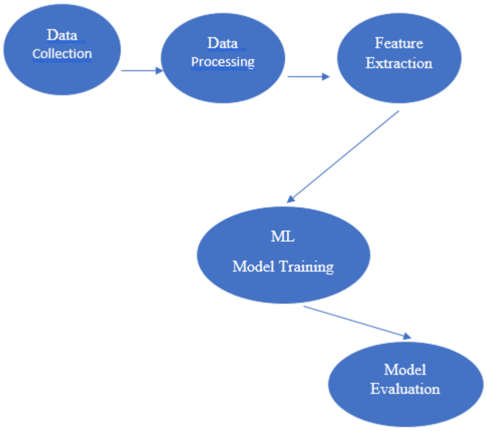

Figure 1: The Machine Learning Pipeline for MS Sclerosis

4.1 Data Preprocessing

We began by looking at how our data is organized. We found that two variables, called ”Initial_EDSS” and ”Final_EDSS,” were missing information in 148 instances. Because of this, we decided to remove these variables from our dataset. TABLE 1 provides an overview of what our dataset looks like.

4.2 Data Exploration

We used visualizations to check how the dependent variable is spread across different independent variables, using a stacked bar chart. When looking at the distribution of the dependent variable (FIGURE 2), we found a balance (CDMS=125, non-CDMS=146). However, when examining how gender relates to the dependent variable “group”, we noticed an uneven distribution. This suggests a possible connection between gender and the dependent variable “group” (FIGURE 3).

Similarly, the distribution of breastfeeding across the dependent variable “group” showed an uneven pattern, indicating a potential link between breastfeeding and the dependent variable “group” (FIGURE 4). This same trend was observed in other variables like Varicella, Oligoclonal, LLSSEP, ULSSEP, VEP, BAEP, Cortical_MRI, Periventricular MRI, and Spinal Cord MRI. These observations shown in FIGURE 5 - FIGURE 15, suggest there might be a relationship between these variables and the dependent variable “group”.

Table 1: Data Structure

Variable	Non-Null
Age	273
Schooling	272
Breastfeeding	273
Varicella	273
Initial_Symptom	272
Mono_or_Polysymptomatic	273
Oligoclonal_Bands	273
LLSSEP	273
ULSSEP	273
VEP	273
BAEP	273
Periventricular_MRI	273
Cortical_MRI	273
Infratentorial_MRI	273
Spinal_Cord_MRI	273
Initial_EDSS	125
Final_EDSS	125
group	273

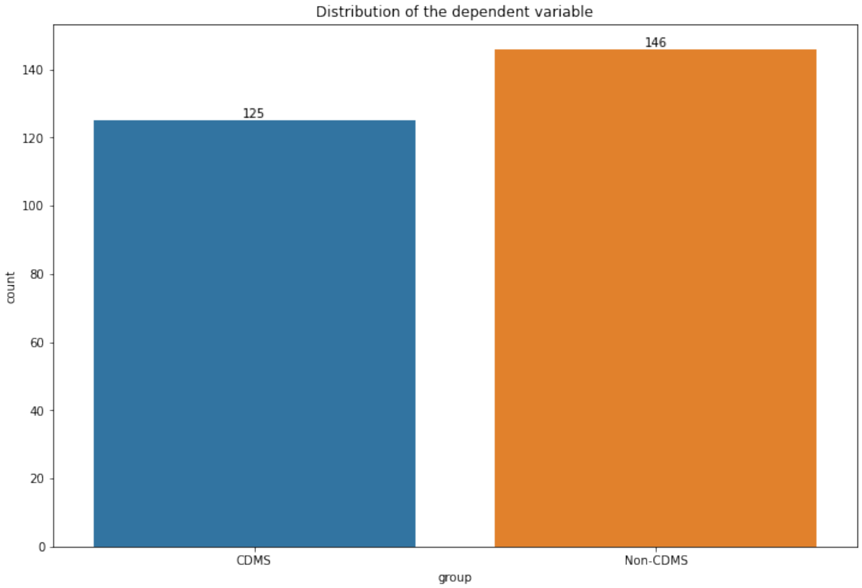

Figure 2: Distribution of the dependent variable

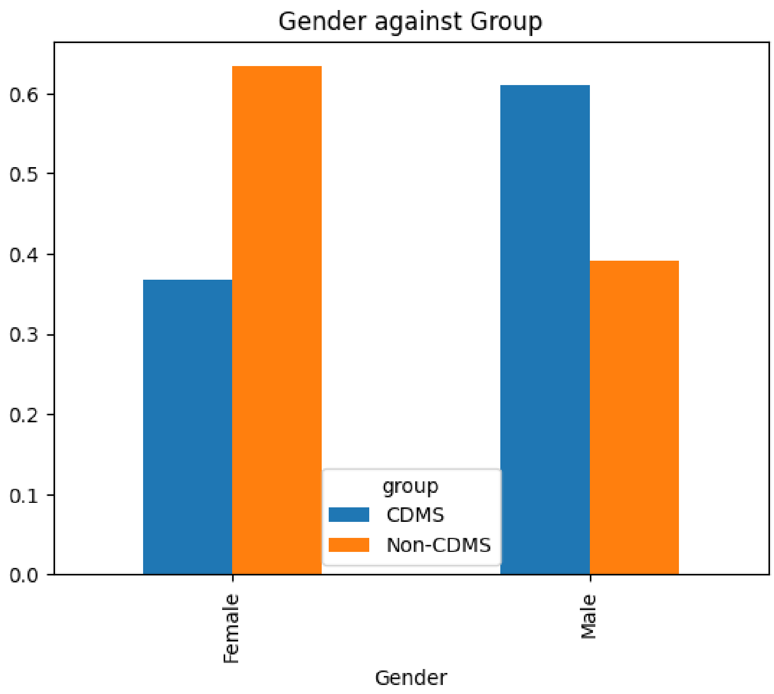

Figure 3: Distribution of the Gender against the dependent variable (Group)

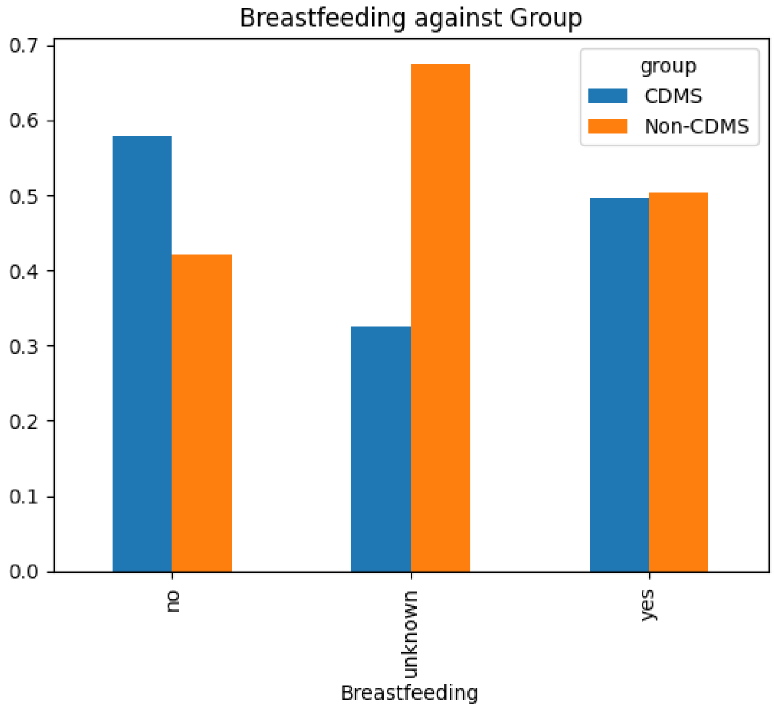

Figure 4: Distribution of the Breastfeeding against the dependent variable (Group)

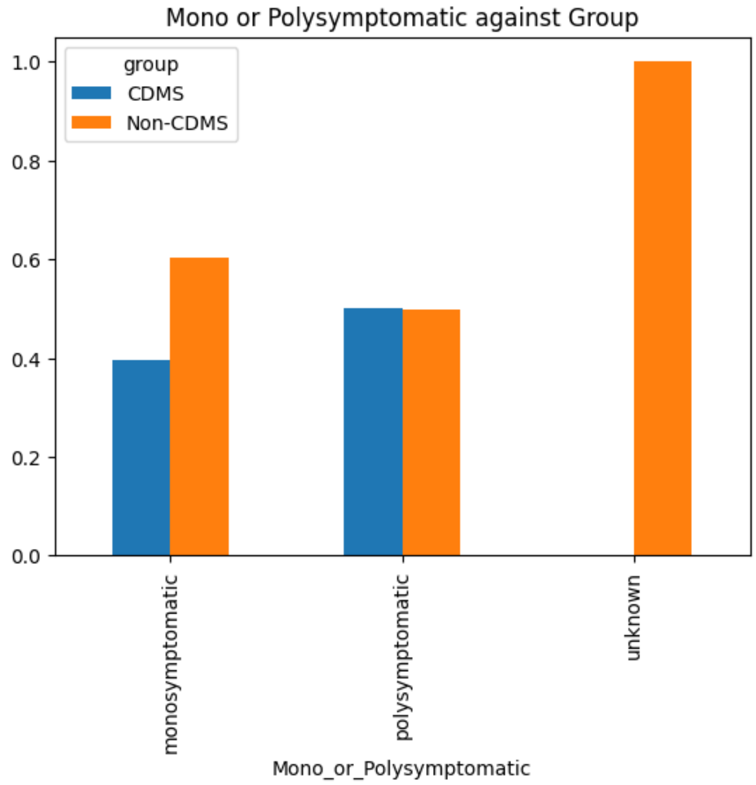

Figure 5: Distribution of the Mono or Polysymptomatic against the dependent variable (Group)

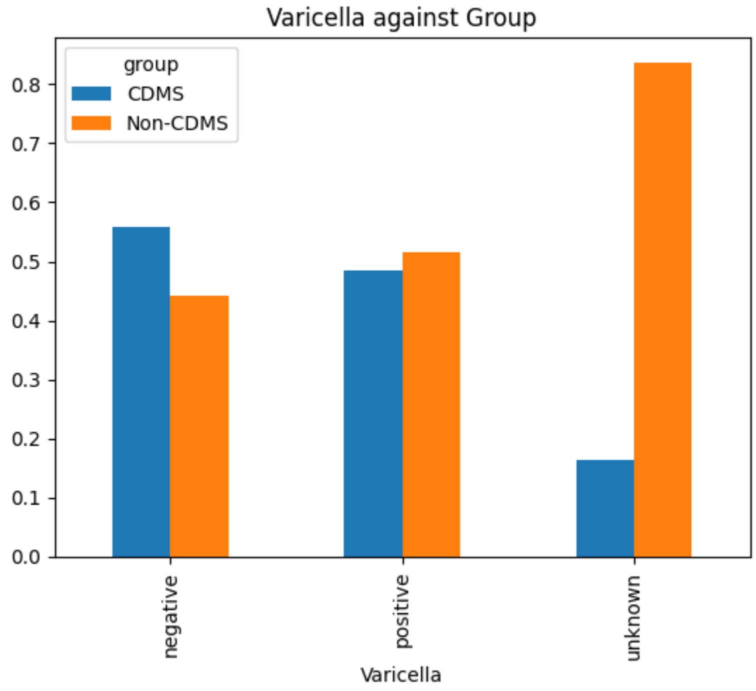

Figure 6: Distribution of the Varicella against the dependent variable (Group)

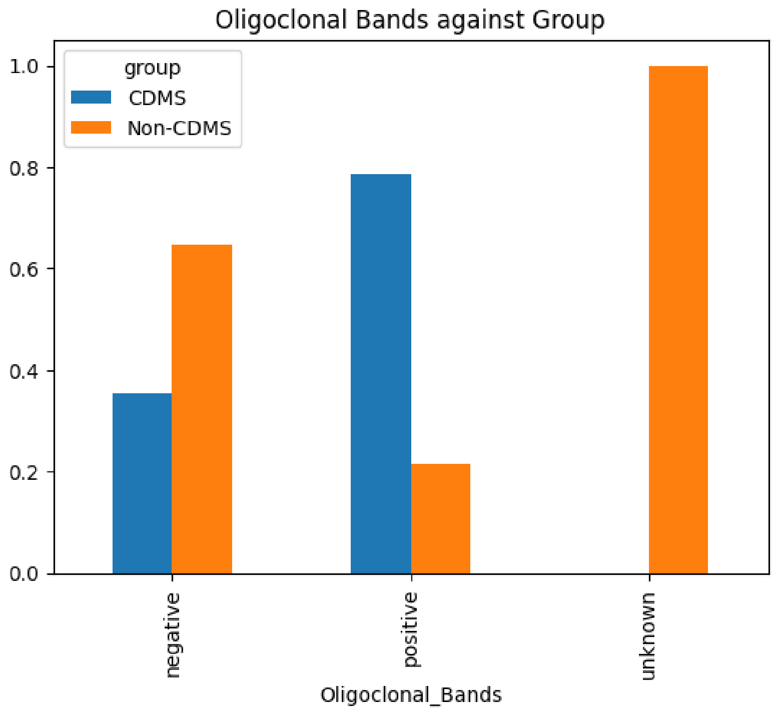

Figure 7: Distribution of the Oligoclonal Bands against the dependent variable (Group)

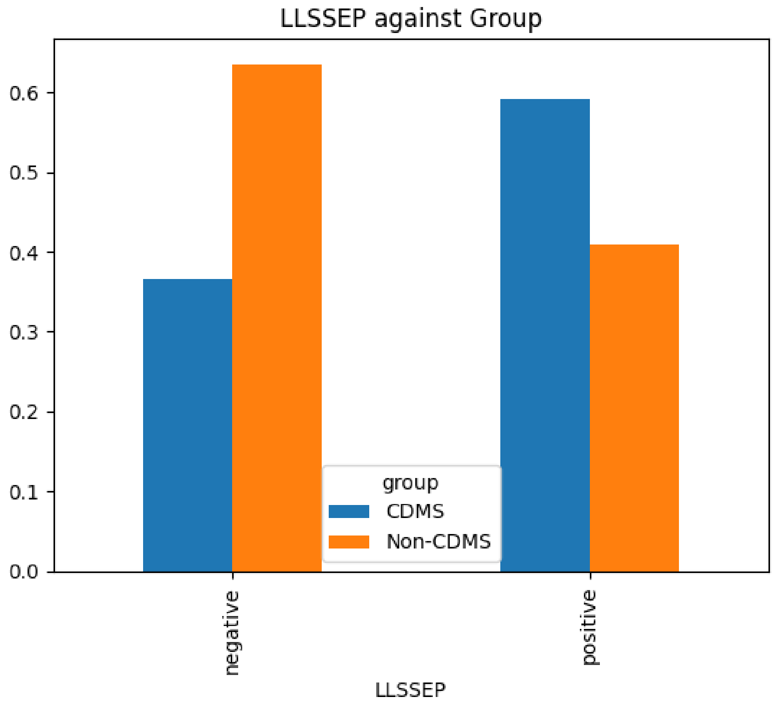

Figure 8: Distribution of the LLSSEP against the dependent variable (Group)

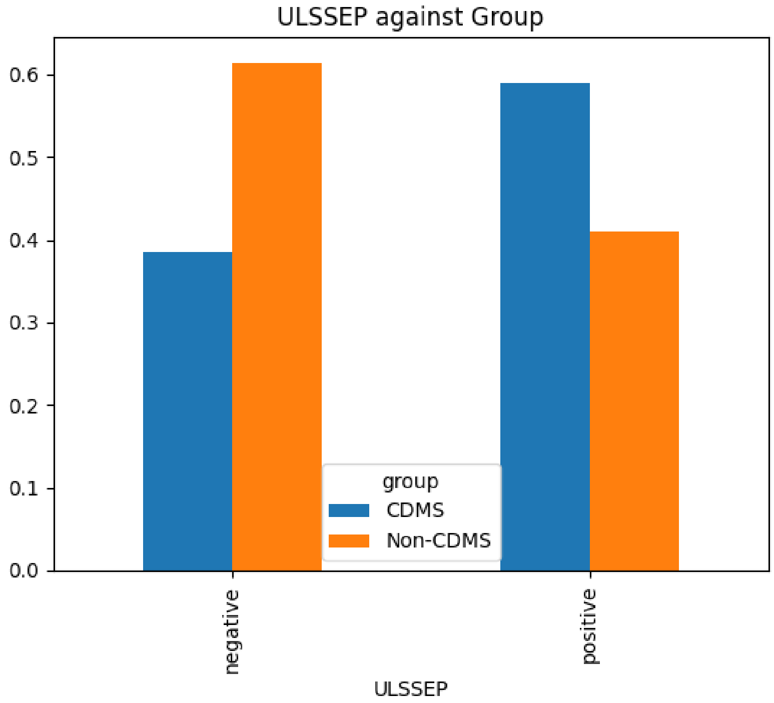

Figure 9: Distribution of the ULSSEP against the dependent variable (Group)

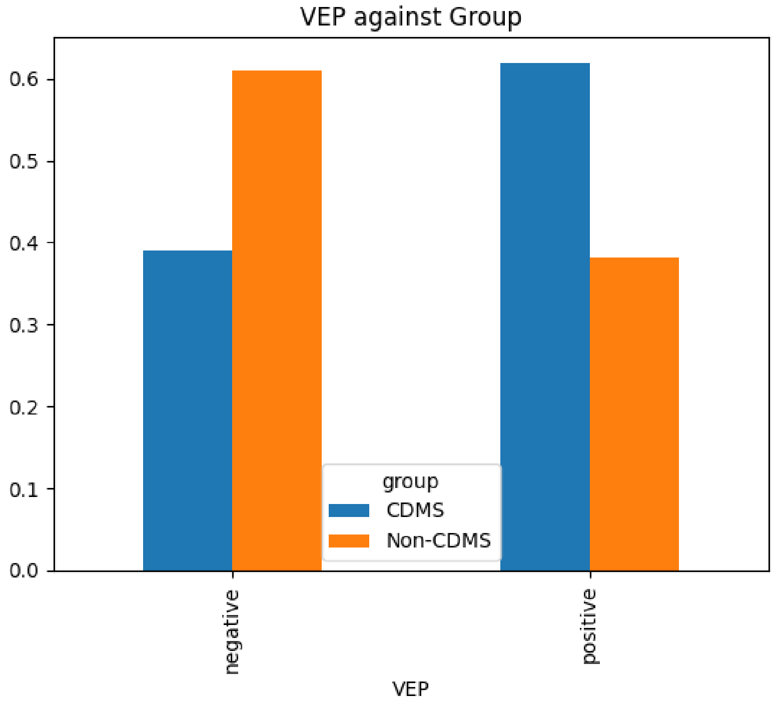

Figure 10: Distribution of the VEP against the dependent variable (Group)

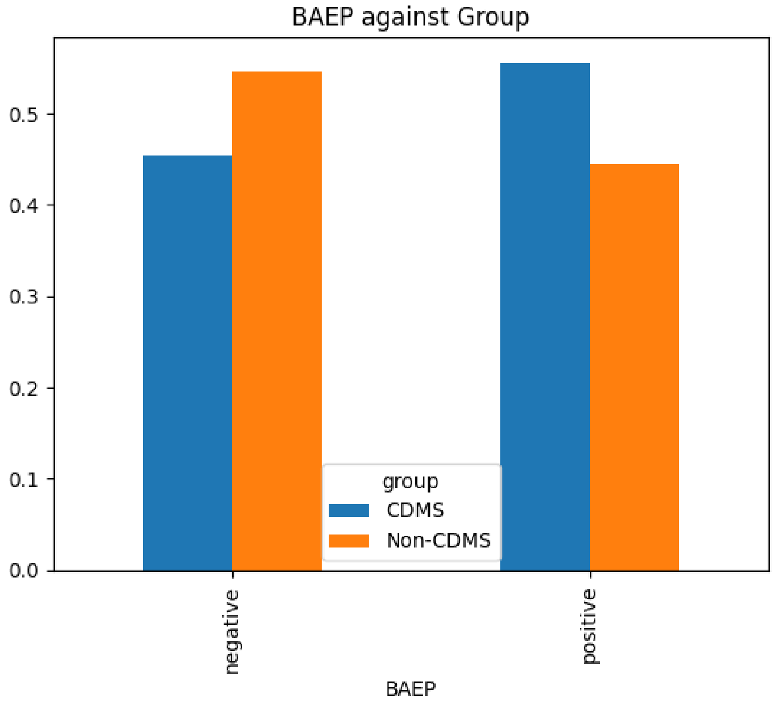

Figure 11: Distribution of the BAEP against the dependent variable (Group)

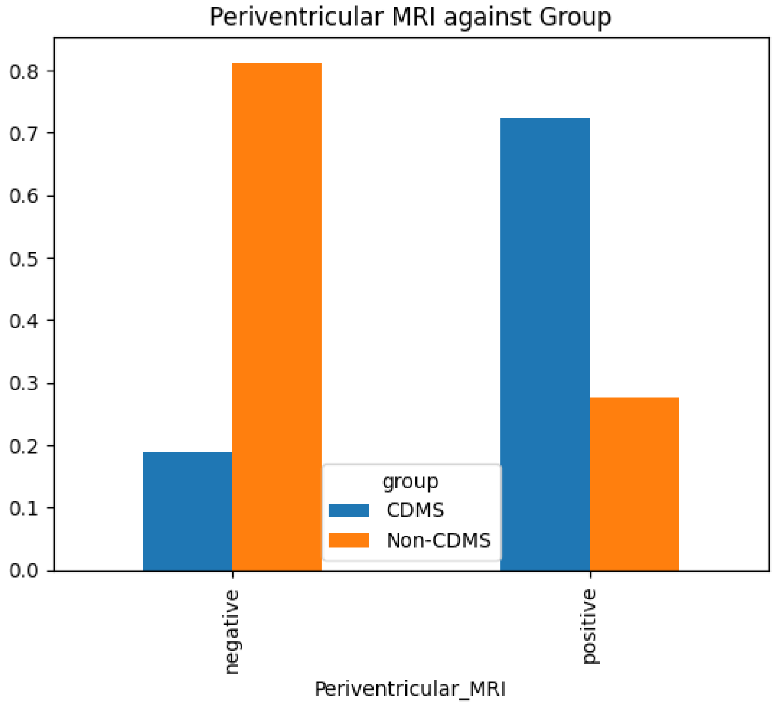

Figure 12: Distribution of the Periventricular MRI against the dependent variable (Group)

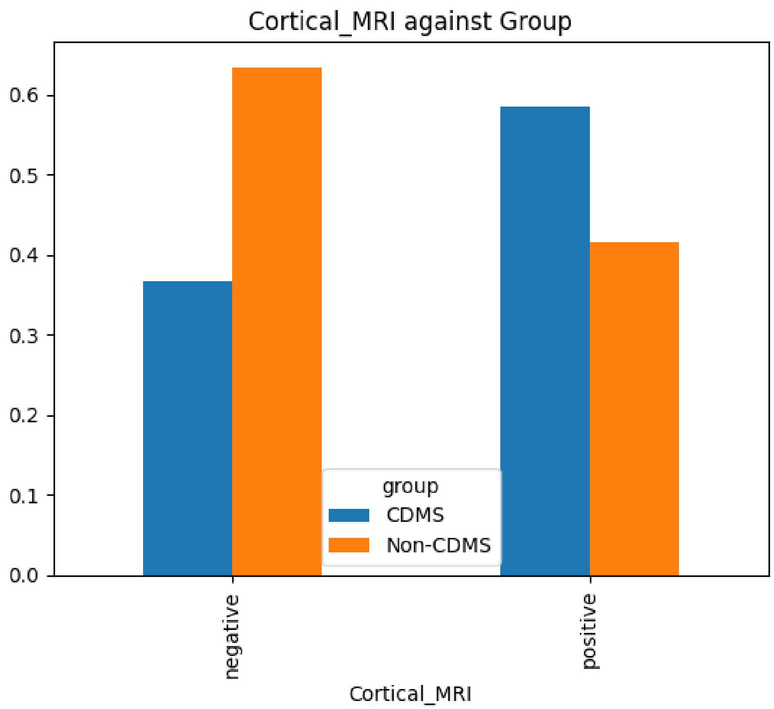

Figure 13: Distribution of the Cortical MRI against the dependent variable (Group)

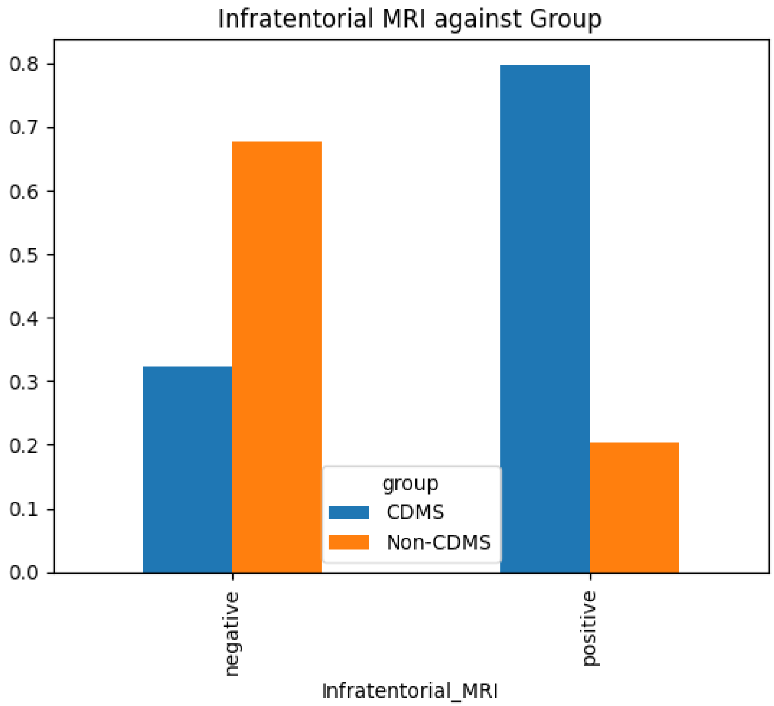

Figure 14: Distribution of the Infratentorial MRI against the dependent variable (Group)

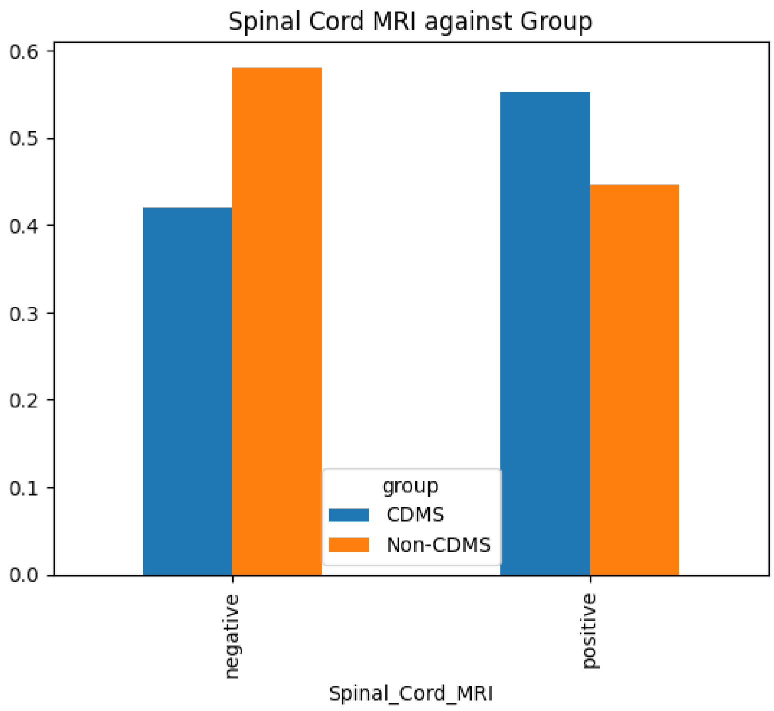

Figure 15: Distribution of the Spinal Cord MRI against the dependent variable (Group)

4.3 Hyper-Parameter Tuning

To find the optimal parameter values for our machine learning models, we conducted hyperparameter tuning. This process involves systematically adjusting the configuration settings (hyperparameters) of a machine learning model to improve its performance. After the tuning process, we identified the optimal parameter values for our models, and you can see them listed below in TABLE 2-TABLE 4.

Table 2: Hyper-Parameter Tuning Random Forest

Model	Bootstrap	Max Features	Estimators	OOB-score
Random Forest	True	8	128	True

Table 3: Hyper-Parameter Tuning Xgboost

Model	Learning Rate	Max Depth	Estimators
Xgboost	0.15	5	150

For the Random Forest, the ”Bootstrap” parameter was set to ”True,” which means the model randomly sampled data with replacement during training. ”Max Features” was set to 8, indicating the maximum number of features considered for splitting a node. The ”Estimators” parameter was

Table 4: Hyper-Parameter Tuning Decision Tree

Model	Min Sample Split	Max Depth
Decision Tree	2	3

set to 128, representing the number of decision trees in the Random Forest. Lastly, the ”OOB-score” was set to ”True,” which means we used out-of-bag samples to estimate the model’s performance.

For the XGBoost model, the ”Learning Rate” is like the step size during learning – the optimal value is 0.15, a moderate value. ”Max Depth” limits how deep the decision trees can go; here, which is 5, preventing overly complex trees. Finally, ”Estimators” represent the number of decision trees in the XGBoost, 150 trees were optimal to capture a diverse range of patterns.

For the Decision Tree, the ”Min Sample Split” is set to 2, meaning that the tree will continue to split a node into smaller branches as long as there are at least 2 data points in the node. The ”Max Depth” is set to 3, indicating the maximum number of levels or branches the tree can have. So, the tree stops growing and making decisions after reaching a depth of 3.

4.4 Model Fitting

Using the optimal parameter values, we proceed to fit our machine learning models. The results from the random forest model indicate good performance. The AUC (Area Under the Curve), which measures the model’s ability to distinguish between CDMS and non-CDMS cases, is 0.93 (FIGURE 16), suggesting strong predictive power. The precision for CDMS is 0.83, meaning that when the model predicts CDMS, it is correct about 83% of the time. For non-CDMS, the precision is 0.89. The recall, which represents the model’s ability to identify all actual CDMS cases, is 0.86 for CDMS and 0.87 for non-CDMS. The accuracy which measures the model’s correctness, is 87%. These metrics in TABLE 5, collectively suggest that the model is performing well in predicting CDMS and non-CDMS cases.

Table 5: Performance of the Random Forest Model

	precision	recall	f1-score	support
CDMS	0.83	0.86	0.85	35
Non-CDMS	0.89	0.87	0.88	47
accuracy			0.87	82
macro avg	0.86	0.86	0.86	82
weighted avg	0.87	0.87	0.87	82

The results of the Decision Tree model in Table 6, reveals its performance metrics. The AUC (Area Under the Curve), which gauges the model’s ability to distinguish between different cases, is 0.86 (FIGURE 17). Precision for CDMS is 0.76, indicating that when the model predicts CDMS, it is correct about 76% of the time. For non-CDMS, the precision is higher at 0.86. The recall, which measures the model’s ability to find all actual CDMS cases, is 0.83 for CDMS and 0.81 for non-

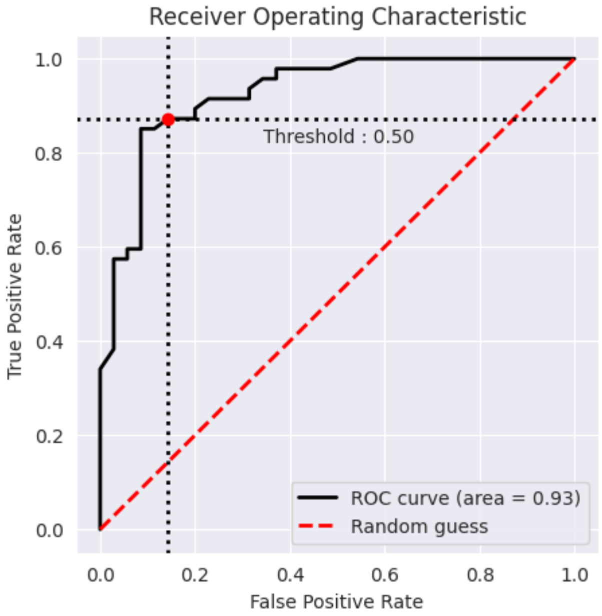

Figure 16: ROC for Random Forest

CDMS. The model is about 83% effective in identifying actual CDMS cases and 81% effective for non-CDMS. The accuracy of the model, reflecting its correctness across all predictions, is 82%.

Table 6: Performance of the Decision Tree Model

	precision	recall	f1-score	support
CDMS	0.76	0.83	0.79	35
Non-CDMS	0.86	0.81	0.84	47
accuracy			0.82	82
macro avg	0.81	0.82	0.81	82
weighted avg	0.82	0.82	0.82	82

The outcome of the Xgboost model in TABLE 7, indicates its effectiveness in predicting whether a patient has Clinically Definite MS (CDMS) or not. The AUC (Area Under the Curve) is 0.90 (FIGURE 18), suggesting strong overall performance. The precision for CDMS is 0.75, meaning that when the model predicts CDMS, it is correct about 75% of the time. For non-CDMS, the precision is higher at 0.83. The recall, which measures the model's ability to find all actual CDMS cases, is 0.77 for CDMS and 0.81 for non-CDMS. The model is about 77% effective in identifying actual CDMS cases and 81% effective for non-CDMS. The accuracy of the model, reflecting its correctness across all predictions, is 79%.

The outcome of the logistic regression model is summarized by several performance metrics. The AUC (Area Under the Curve), which measures the model's ability to distinguish between different cases, is 0.90 (FIGURE 19). Precision for CDMS is 0.84, meaning that when the model predicts CDMS, it is correct about 84% of the time. For non-CDMS, the precision is 0.82. The recall,

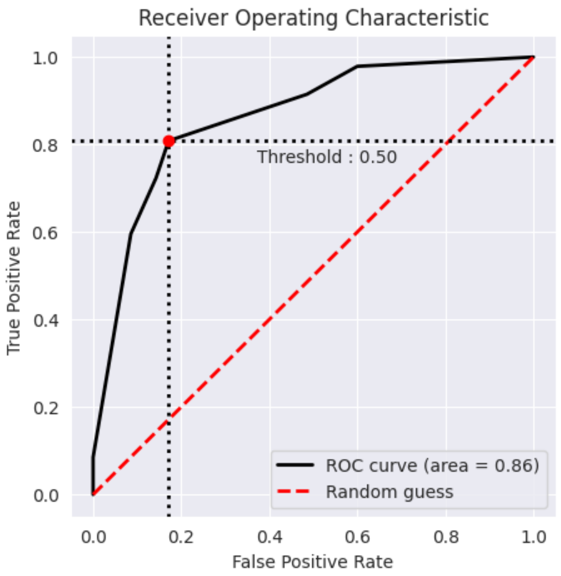

Figure 17: ROC for Decision Tree

Table 7: Performance of the Xgboost Model

	precision	recall	f1-score	support
CDMS	0.75	0.77	0.76	35
Non-CDMS	0.83	0.81	0.82	47
accuracy			0.79	82
macro avg	0.79	0.79	0.79	82
weighted avg	0.79	0.79	0.79	82

which shows the model's ability to identify all actual CDMS cases, is 0.74 for CDMS and 0.89 for non-CDMS. The model is about 74% effective in capturing actual CDMS cases and 89% effective for non-CDMS. The Accuracy of the model, reflecting its correctness across all predictions, is 83%. These metrics in TABLE 8, suggest that the logistic regression model is performing well in predicting both CDMS and non-CDMS cases.

Table 8: Performance of the Logistic Regression Model

	precision	recall	f1-score	support
CDMS	0.84	0.74	0.79	35
Non-CDMS	0.82	0.89	0.86	47
accuracy			0.83	82
macro avg	0.83	0.82	0.82	82
weighted avg	0.83	0.83	0.83	82

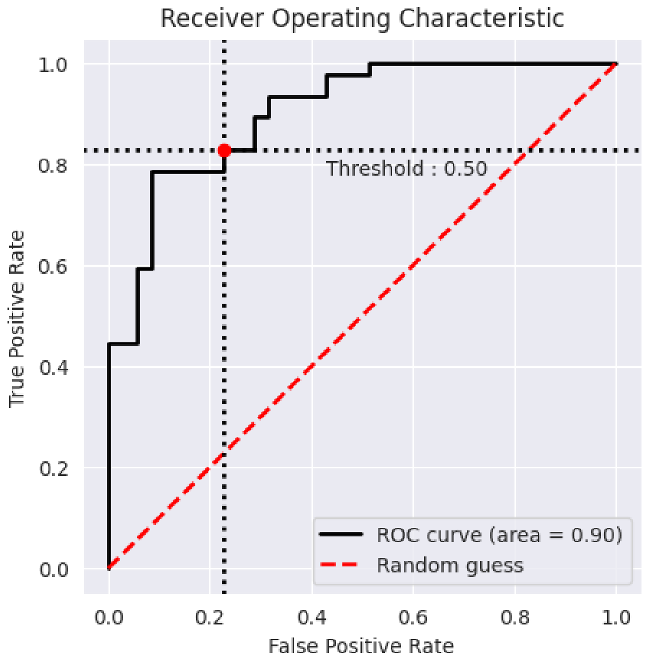

Figure 18: ROC for Xgboost Model

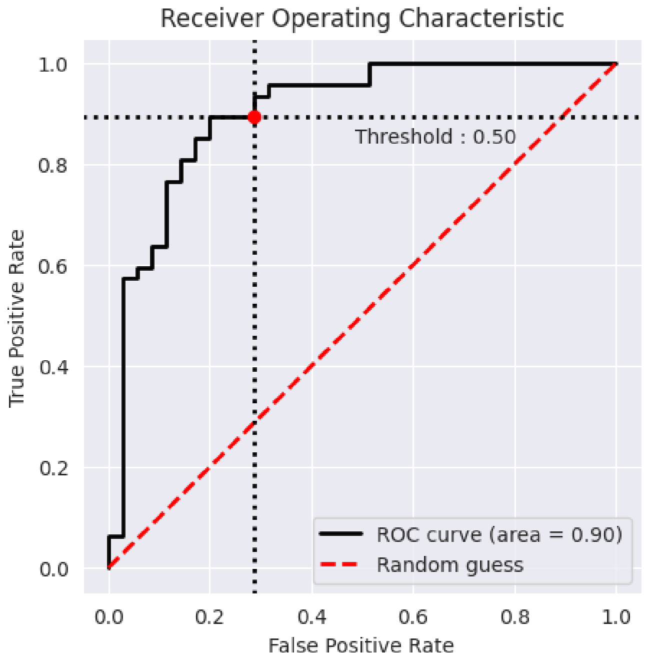

Figure 19: ROC for Logistic Model

In comparison with the other models, the random forest model performs better in predicting both CDMS and non-CDMS instances, according to the performance metrics given. We propose the

Random Forest as the best model for this application because of its superior overall performance. As a result of this advice, we created a variable significance plot and looked into the variables that have the most impact on CDMS prediction further.

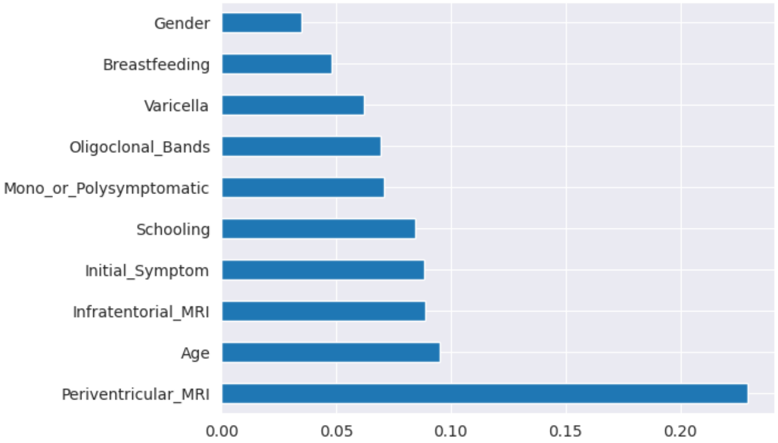

Figure 20: Variable Importance Plot

As illustrated in FIGURE 20, the variable importance plot reveals the top 10 factors that strongly influence the prediction of CDMS. These influential factors, listed in order of importance, are Periventricular MRI, Age, Infratentorial MRI, Initial Symptom, Schooling, Mono or Polysymptomatic, Oligoclonal Bands, Varicella, Breastfeeding, and Gender. This means that these particular aspects play a significant role in determining whether a patient is likely to have Clinically Definite MS (CDMS) based on the model's analysis.

5. CONCLUSION

In this study, machine learning algorithms were employed to predict the progression of Clinically Isolated Syndrome (CIS) to Clinically Definite MS (CDMS) using open-source data. The selected algorithms—Random Forest, XGBoost, Decision Tree, and Logistic Regression—were carefully tuned and evaluated for their performance.

Random Forest model demonstrated superior overall performance, with an AUC of 0.93, accuracy of 87%, and balanced precision and recall for both CDMS and non-CDMS cases. XGBoost, Decision Tree, and Logistic Regression models also exhibited commendable predictive capabilities, with AUCs ranging from 0.86 to 0.90 and accuracies from 79% to 83%.

Variables such as Periventricular MRI, Age, Infratentorial MRI, Initial Symptom, Schooling, Mono or Polysymptomatic, Oligoclonal Bands, Varicella, Breastfeeding, and Gender were identified as the top 10 influential factors in predicting CDMS progression.

The findings underscore the potential of machine learning in enhancing the early prognosis of MS progression, providing valuable insights for timely therapeutic interventions. The Random Forest model, in particular, stands out as a robust tool for predicting CDMS, offering clinicians a reliable means to identify patients at risk. This research contributes to the evolving landscape of predictive analytics in neurological disorders, showcasing the effectiveness of machine learning in leveraging diverse clinical and demographic variables. Future studies may explore the integration of additional data sources and advanced feature engineering techniques to further refine predictive models. The application of machine learning in predicting MS progression holds promise for personalized medicine, enabling tailored interventions based on individual patient profiles.

The study is based on a specific cohort of Mexican mestizo patients, and generalizability to other populations may require additional validation. The predictive models are based on retrospective data, and prospective studies are warranted to validate their real-world applicability.

In conclusion, this research marks a significant step towards harnessing machine learning for early prediction of MS progression, offering clinicians a valuable tool to improve patient care and contribute to the evolving field of predictive healthcare analytics.

References

[1] Lassmann H. Multiple sclerosis pathology. Cold Spring Harb Perspect Med. 2018;8:a028936.

[2] Thompson AJ, Baranzini SE, Geurts J, Hemmer B, Ciccarelli O. Multiple sclerosis. Lancet. 2018;391:1622-1636.

[3] Rodríguez Murúa S, Farez MF, Quintana FJ. The Immune Response in Multiple Sclerosis. Annu Rev Pathol. 2022;17:121-139.

[4] Zhang H, Alberts E, Pongratz V, Mühlau M, Zimmer C, et al. Predicting Conversion From Clinically Isolated Syndrome to Multiple Sclerosis–an Imaging-Based Machine Learning Approach. NeuroImage Clin. 2019;21:101593.

[5] Walton C, King R, Rechtman L, Kaye W, Leray E, et al. Rising Prevalence of Multiple Sclerosis Worldwide: Insights From the Atlas of MS, Third Edition. Mult Scler.2020;26:1816-1821.

[6] Vollmer TL, Nair KV, Williams IM, Alvarez E. Multiple Sclerosis Phenotypes as a Continuum: The Role of Neurologic Reserve. Neurol Clin Pract. 2021;11:342-351.

[7] Terranova N, Venkatakrishnan K, Benincosa LJ. Application of Machine Learning in Translational Medicine: Current Status and Future Opportunities. AAPS J. 2021;23:74.

[8] Hutchinson L, Steiert B, Soubret A, Wagg J, Phipps A, et al. Models and Machines: How Deep Learning Will Take Clinical Pharmacology to the Next Level. CPT Pharmacometrics Syst Pharmacol. Mar. 2019;8:131-134.

[9] Cavaliere C, Vilades E, Alonso-Rodríguez MC, Rodrigo MJ, Pablo LE, et al. Computer-Aided Diagnosis of Multiple Sclerosis Using a Support Vector Machine and Optical Coherence Tomography Features. Sensors Basel. 2019;19:5323.

[10] Seccia R, Gammelli D, Dominici F, Romano S, Landi AC, et al. Considering Patient Clinical History Impacts Performance of Machine Learning Models in Predicting Course of Multiple Sclerosis. PLoS One. Mar. 2020;15:e0230219.

[11] Eshaghi A, Young AL, Wijeratne PA, Prados F, Arnold DL, et al. Identifying Multiple Sclerosis Subtypes Using Unsupervised Machine Learning and MRI Data. Nat Commun. 2021;12:2071.

[12] Lancet T. ICD-11: A Brave Attempt at Classifying a New World. Lancet. 2018;391:2476.

[13] Briganti G, Le Moine OL. Artificial Intelligence in Medicine: Today and Tomorrow. Front Med. 2020;7:27.

[14] Dobson R, Giovannoni G. Multiple Sclerosis - A Review. Eur J Neurol. 2019;26:27-40.

[15] Damasceno A, Pimentel-Silva LR, Damasceno BP, Cendes F. Exploring the Performance of Outcome Measures in MS for Predicting Cognitive and Clinical Progression in the Following Years. Mult Scler Relat Disord. 2020;46:102513.

[16] Tousignant A, Lemaître P, Precup D, Arnold DL, Arbel T. Prediction of Disease Progression in Multiple Sclerosis Patients Using Deep Learning Analysis of MRI Data. In International conference on medical imaging with deep learning. Proceedings of the Machine Learning Research. 2019;102:483–492.

[17] Brichetto G, Monti Bragadin M, Fiorini S, Battaglia MA, Konrad G, et al. The Hidden Information in Patient-Reported Outcomes and Clinician-Assessed Outcomes: Multiple Sclerosis as a Proof of Concept of a Machine Learning Approach. Neurol Sci. 2020;41:459-462

[18] Raeisi Z, Ramezannezad P, Ahmadzade M, Tarahomi S. Analyzing Clinical Symptoms in Multiple Sclerosis Using Data Mining. Tehran University of Medical Sciences Journal. 2017;75:39-48

[19] Storelli L, Azzimonti M, Gueye M, Vizzino C, Preziosa P, Tedeschi G et al. A Deep Learning Approach to Predicting Disease Progression in Multiple Sclerosis Using Magnetic Resonance Imaging. Invest Radiol, 2022;57:423-432.

[20] Vázquez-Marrufo M, Sarrias-Arrabal E, García-Torres M, Martín-Clemente R, Izquierdo G. A Systematic Review of the Application of Machine-Learning Algorithms in Multiple Sclerosis. Neurol 2022;38.

[21] Zhao Y, Wang T, Bove R, Cree B, Henry R, Lokhande H et al. Ensemble Learning Predicts Multiple Sclerosis Disease Course in the Summit Study. NPJ Digit Med. 2020;3:135.

[22] Law MT, Traboulsee AL, Li DK, Carruthers RL, Freedman MS, et al. Machine Learning in Secondary Progressive Multiple Sclerosis: An Improved Predictive Model for Short-Term Disability Progression. Mult Scler J Exp Transl Clin. 2019;5:2055217319885983.

[23] Andorra M, Freire A, Zubizarreta I, de Rosbo NK, Bos SD et al., Predicting Disease Severity in Multiple Sclerosis Using Multimodal Data and Machine Learning. J Neurol. 2023:1-7

[24] Xia Z, Secor E, Chibnik LB, Bove RM, Cheng S, et al. Modeling Disease Severity in Multiple Sclerosis Using Electronic Health Records. PLOS ONE. 2013;8:e78927.

[25] Wang SH, Tang C, Sun J, Yang J, Huang C, et al. Multiple Sclerosis Identification by 14-Layer Convolutional Neural Network With Batch Normalization, Dropout, and Stochastic Pooling. Front Neurosci. 2018;12:818.

[26] Zhao Y, Healy BC, Rotstein D, Guttmann CR, Bakshi R, Weiner HL et al. Exploration of Machine Learning Techniques in Predicting Multiple Sclerosis Disease Course. PLOS ONE. Apr. 2017;12:e0174866.

[27] Plati D, Tripoliti E, Zelilidou S, Vlachos K, Konitsiotis S. Multiple Sclerosis Severity Estimation and Progression Prediction Based on Machine Learning Techniques. Int Conf IEEE Eng Med Biol Soc. 2022:1109-1112.

[28] Ramanujam R, Zhu F, Fink K, Karrenbauer VD, Lorscheider J, Benkert P et al. Accurate Classification of Secondary Progression in Multiple Sclerosis Using a Decision Tree. Mult Scler J. 2021;27:1240-1249.