의학박사 학위 논문

Machine Learning on Brain MRI data for Differential diagnosis of Multiple sclerosis and Neuromyelitis optica spectrum disorder

다발경화증과 시신경척수염범주질환의 감별진단을 위한

뇌 자기공명영상에 대한 기계학습

울 산 대 학 교 대 학 원

의 학 과

김 현 진

Machine Learning on Brain MRI data for Differential diagnosis of Multiple sclerosis and Neuromyelitis optica spectrum disorder

지 도 교 수　김 광 국 & 강 동 화

이 논문을 의학박사 학위 논문으로 제출함

2018 년 12 월

울 산 대 학 교　대 학 원

의 학 과

김 현 진

김현진의 의학박사 학위 논문을 인준함

심 사 위 원　김 광 국　(인)

심 사 위 원　강 동 화　(인)

심 사 위 원　임 영 민　(인)

심 사 위 원　강 봉 희　(인)

심 사 위 원　이 은 재　(인)

울 산 대 학 교　대 학 원

2018 년 12 월

Abstract

Introduction: Distinguishing between neuromyelitis optica spectrum disorder (NMOSD) and multiple sclerosis (MS) is important because their treatments differ, and disease-modifying treatments for MS can worsen NMOSD. Brain magnetic resonance imaging (MRI) is one of the most important diagnostic tools used to differentiate between the two diseases. To date, considerable effort has been put in the identification of the brain MRI characteristics that enable the differentiation between MS and NMOSD. Machine learning has been studied as a method for identifying medical images. The study aimed to implement a supervised machine-learning method to perform differential diagnosis of MS and NMOSD by using brain MRIs.

Methods: Fluid-attenuated inversion recovery (FLAIR) MRIs were acquired from patients with relapsing-remitting MS (RRMS) and NMOSD with aquaporin-4 immunoglobulin G (AQP4-IgG) admitted at the Asan Medical Center, Seoul, Korea, between 2005 and 2017. FLAIR MRIs were used for a machine-learning method based on the combination of lesion frequency analysis for feature voxel selection and support vector machines (SVM) for classification algorithm. Diagnostic performance of machine learning was compared to that of two neurologists with more than two years of clinical experience in demyelinating disease.

Results: Final analysis included 746 and 292 MRIs from 172 patients with RRMS and 97 patients with NMOSD with AQP4-IgG, respectively. Lesion frequency

analysis found that lesions adjacent to the lateral ventricle and in the inferior temporal lobe were frequently observed in RRMS, and dorsal medulla, cerebral peduncle/internal capsule, and corpus callosum lesions were frequently observed in NMOSD. The performance of SVM was 57.5% sensitivity, 78.4% specificity, and 63.3% accuracy, which showed a fair level of agreement with human raters (Cohen's κ, rater A=0.279, rater B=0.262).

Conclusion: Machine learning using brain MRI data could discern RRMS and NMOSD with comparable accuracy to that of clinicians, encouraging the application of machine learning-aided diagnosis in clinical practice.

Key words: brain MRI, multiple sclerosis, neuromyelitis optica spectrum disorder, machine learning, support vector machine

Contents

List of tables & figures

Introduction

1. **Diagnosis of NMOSD**

Neuromyelitis optica spectrum disorder (NMOSD) is an inflammatory disease of the central nervous system (CNS), mostly involving the optic nerve and the spinal cord.[1)] Although it is a rare disease, an epidemiological study estimated its prevalence to be as high as 10 per 100,000 in an Afro-Caribbean population,[2)] whereas a recent study reported the prevalence to be 4.1 per 100,000 in Hokkaido, a part of northern Japan, in Asia.[3)] There is no exact data yet on the prevalence of NMOSD in Korea. A recent review suggested a similar global prevalence of NMOSD as 5 or less per 100,000 habitants, compared to the uneven global distribution of multiple sclerosis (MS) prevalence.[4)] NMOSD characteristically demonstrates a high female predominance of approximately three to nine females affected to every one male affected,[5)] most with a disease-specific autoantibody to aquaporin-4 (AQP4-Ab),[6)] and frequently manifesting as severe bilateral/recurrent optic neuritis or severe longitudinally extensive transverse myelitis (LETM).[7)] However, a number of studies have revealed that brain abnormalities are not rare in NMOSD. The incidence of reported brain abnormalities in NMOSD was approximately 59% to 79%.[8, 9)] Various diseases such as MS, inflammatory diseases, vascular diseases, infection, or malignancy can mimic NMOSD by either involving optic nerves and/or spinal cords, manifesting bilateral optic neuritis or LETM,[10)] and/or showing brain lesions resembling those of NMOSD. Therefore, the diagnosis of NMOSD can be complicated.

The diagnostic criteria of neuromyelitis optica (NMO) stemmed from the original criteria in 1999,[11)] through those revised in 2006,[12)] and finally to the first international consensus criteria in 2015.[1)] The latest criteria have adopted the broader term of NMOSD to include

patients with limited manifestations. Further, NMOSD has been classified into two types according to the new diagnostic criteria: NMOSD with aquaporin-4 immunoglobulin G (AQP4-IgG) and NMOSD without AQP4-IgG or with unknown AQP4-IgG status. NMOSD with AQP4-IgG refers to patients who have at least one core clinical characteristics of NMOSD in optic nerve, spinal cord, dorsal medullar, brainstem, diencephalon, or cerebrum; who test positive for AQP4-IgG; and in whom alternative diagnoses are excluded.[1)]

According to the 2015 criteria, the presence of AQP4-Ab is crucial for the diagnosis of NMOSD with AQP4-IgG. Nevertheless, clinical and radiological differential diagnosis of NMOSD with AQP4-IgG remains important for the following reasons: (1) in clinical practice, the AQP4-Ab assay may not be performed or may not be readily available for all patients with inflammatory disease of the CNS. Rather, clinicians need to identify patients with probable NMOSD in whom the AQP4-Ab assay should be performed; (2) the test result for AQP4-Ab could be affected by factors such as test methods and clinical (relapse or remission) situations; (3) many diseases, including inflammatory, infectious, or neoplastic conditions, can involve the CNS and mimic the clinical and radiological features of NMOSD; and (4) some patients with NMOSD do not have AQP4-Ab (NMOSD without AQP4-IgG or with unknown AQP4-IgG status).[1)]

2. **Differential diagnosis of MS and NMOSD**

Both MS and NMO are inflammatory diseases of the CNS with relapsing courses. As they share some clinical and radiological features, there has been considerable debate on whether these two diseases are fundamentally different. However, since the discovery of AQP4-Ab, the disease-specific autoantibody to NMOSD, studies have confirmed that the two diseases have distinct features in their epidemiology, serology, pathology, response to treatment, and

prognosis.[13] Differentiating NMOSD from MS is important, because treatment for each disease differs, and disease-modifying treatments for MS, including interferon-β, fingolimod, and natalizumab can worsen NMOSD.[14-16]

3. Role of brain MRI for differential diagnosis of two diseases

Due to the presence of NMOSD without AQP4-IgG and the fact that approximately 5% to 42% of patients with NMO fulfill the Barkhof magnetic resonance imaging (MRI) criteria for MS,[17, 18] brain MRI is one of the most important diagnostic tools that can enable the differentiation between the two diseases. To date, there has been considerable effort to identify brain MRI characteristics that differ between MS and NMOSD.[17, 19, 20] A recent study identified the criteria of "at least one lesion adjacent to the body of the lateral ventricle and in the inferior temporal lobe; or the presence of a subcortical U-fiber lesion; or a Dawson's finger-type lesion," which could distinguish patients with MS from those with NMOSD with 92% sensitivity and 96% specificity.[17] In addition, the same criteria were used to distinguish MS from myelin oligodendrocyte glycoprotein (MOG)-antibody disease with 90.9% sensitivity and 95.2% specificity.[20] However, the abovementioned studies excluded brain MRIs without abnormalities, and the brain lesion distribution criteria was difficult to apply for clinicians who had minimal experience with demyelinating disease. Therefore, recent efforts have focused on alternative approaches for the analysis of neuroimaging data.

4. Machine learning

Machine-learning techniques are based on algorithms that can automatically extract information from brain images and classify individual structural or functional brain images by maximizing the distance between groups of images.[21] Machine learning is typically

classified into supervised and unsupervised learning.[22, 23] Supervised machine learning uses a training dataset labeled by humans to define known answers. It may expedite classification or regression processes with large datasets and can be useful for predicting clinical outcomes or classifying clinical diagnoses. However, it requires a human labeling process, which is often tedious and time-consuming. Examples of supervised learning methods include the support vector machine (SVM), decision tree, linear regression, logistic regression, naive Bayes, and random forest.[23] In contrast, unsupervised machine learning does not use human-defined answers. Instead, it seeks to identify hidden patterns in large datasets, which cannot be usually recognized by humans. Therefore, unsupervised learning may be useful in seeking novel disease mechanisms, genotypes, and phenotypes. Examples of unsupervised learning include K-means, mean shift, affinity propagation, hierarchical clustering, and Gaussian mixture modeling.[24]

In machine learning, variables used as input data are generally referred to as features, which may be numerical or nominal values. Because the performance of machines is variable according to the entered features, it is very important to select and extract features from the data appropriately. The entered features are usually determined by researchers and data scientists. Various feature selection methods have been developed to enhance the selection process and establish machine models with high accuracy. Using these features, machine-learning algorithms determine the optimal decision boundary, select features and develop a model, and conduct a set task. For imaging data, various image features such as the size, location, shape, and signal intensities of the lesion can be used for machine learning.[23]

5. Support vector machine

SVM is a supervised machine-learning method, which is useful for developing a model to classify an object in one category or another. Therefore, SVM is widely used in clinical imaging analysis, which classifies or categorizes a diagnosis. SVM constructs a hyper-plane in a high-dimensional space as the decision surface. To achieve better performance, the margin of separation between classes is maximized. For a nonlinear classification, SVM uses the kernel technique, which implicitly converts the input features into high-dimensional feature spaces. Therefore, selection of the kernel must be appropriate to avoid increases in error rates. Several studies have assessed the diagnostic value of the machine learning (i.e., SVM) for differentiating Parkinson's disease from progressive supranuclear palsy and Alzheimer's disease from elderly controls.[21, 25] Salvatore et al. used an SVM algorithm on T1-weighted brain MRIs in Parkinson's disease and progressive supranuclear palsy, and diagnostic performance showed more than 90% accuracy, sensitivity, and specificity.[21] Magnin et al. applied an SVM algorithm to classify patients with Alzheimer's disease and control subjects with 96.6% mean specificity and 91.5% mean sensitivity.[25]

6. Fluid-attenuated inversion recovery

Fluid-attenuated inversion recovery (FLAIR) is an MRI technique that shows areas of tissue with T2 prolongation as bright while suppressing the image of cerebrospinal fluid (CSF), clearly revealing lesions in proximity to CSF such as juxtacortical and periventricular lesions [26]. FLAIR has been considered superior to T2-weighted images for detecting MS brain lesions, including those in or adjacent to the cerebral cortical gray matter.[27] The evidence-based guidelines from the Magnetic Resonance Imaging in MS (MAGNIMS) network suggested a mandatory MRI sequence: 1) axial proton-density and/or T2-FLAIR/T2-

weighted, 2) sagittal two-dimensional (2D) or three-dimensional (3D) T2-FLAIR, and 3) 2D or 3D contrast-enhanced T1-weighted images.[28)]

7. The aim of the study

The study aimed to implement a supervised machine-learning method (i.e., SVM) able to perform individual differential diagnosis of MS and NMOSD by using brain FLAIR MRIs.

Methods

1. Subjects

We retrospectively reviewed medical records of patients with relapsing-remitting multiple sclerosis (RRMS) and NMOSD who were admitted to the Asan Medical Center, Seoul, Korea, between 2005 and 2017. Patients with RRMS who fulfilled the 2010 McDonald criteria and patients with NMOSD with AQP4-IgG according to the 2015 International Consensus of NMOSD were included in this study. AQP4-IgG testing was performed by a tissue-based indirect immunofluorescence or a cell-based assay. Patients without available brain MRIs were excluded from the study.

2. Data Acquisition

Brain MRI scans including FLAIR sequences were acquired on various 1.5- or 3.0- T scanners (Siemens, GE, Philips) in Asan Medical Center or other centers that referred patients to Asan Medical Center and transferred images.

3. RRMS and NMOSD Classification: Human Raters

Two neurologists (i.e., rater A and B) participated as human raters. Both had one year of experience with demyelinating disease as a fellowship after completion of four years of neurology residency. During the five-year training period, the two neurologists accumulated at least two years of clinical experience diagnosing and treating patients with MS, NMOSD, and other CNS inflammatory diseases. Human raters binarily decided either RRMS or NMOSD based on FLAIR images with the following demographic information: age at time of MRI, age at disease onset, sex, and duration from both first and latest episode of symptoms.

4. RRMS and NMOSD Classification: Machine Learning

4.1. Image Processing

A characteristic of white matter lesion was anomalous hyperintensity in the FLAIR image. To reduce the systemic variation in the FLAIR images due to different MRI equipment and parameters, intensity normalization was applied to each FLAIR image as implemented in the Lesion Segmentation Toolbox (www.statisticalmodelling.de/lst.html) for statistical parametric mapping [29]. In detail, brain area was segmented into gray matter, white matter, and cerebrospinal fluid from the FLAIR image. The voxel intensities of FLAIR image were normalized using the mean intensity of gray matter area as follows (i.e., gray matter belief map in this study):

$$B_v = F_v / \overline{F_{v=\{GM\}}} - \overline{F_{v=\{GM\}}}$$

where B_v and F_v indicated the gray matter belief value, and original intensity of a FLAIR image at a voxel v, respectively; and $\overline{F_{v=\{GM\}}}$ indicated mean of FLAIR intensity in the gray matter voxels.

Threshold for lesion identification from the gray matter belief map was determined based on mean and standard deviation (SD) of concatenated gray matter belief values of brain area across subjects who belonged to training set (see Data Division; mean ± SD of gray matter belief values = 0.0261 ± 0.1199). Estimated threshold for anomalous white matter hyperintensity based on gray matter belief values was 0.32 ($p < 0.01$; mean + SD × 2.54). An individual lesion map was then prepared from the individual gray matter belief map showing which voxel was above the threshold. Spatial normalization procedure was then applied to transform the individual lesion map from the acquisition space to standard Montreal Neurological Institute space with isotropic 3 mm voxel size for further analyses.

4.2. Support Vector Machine for RRMS and NMOSD Classification

SVM is a supervised machine-learning method, which is useful for developing a model to allocate an object to one category or another. In the present study, SVM was used to determine the either NMOSD or RRMS for the inputted FLAIR image. SVM classification was performed using LIBSVM software (www.csie.ntu.edu.tw/~cjlin/libsvm). For the nonlinear SVM classifier, two parameters were specified: C (regularization) and α (parameter for radial basis function kernel). By using a combination of grid search and cross-validation procedures in the training samples, the optimal values of C and α were estimated while considering the generalizability of the trained SVM model. The optimal parameters were determined by grid searching the parameter space and selecting the pair of values (C, α) at which the M-fold cross-validation accuracy was maximum (M = 10 in this study). In order to search for a wide range of values, the values of C and α were varied from 0.125 to 32 in steps of 2 (0.125, 0.25, 0.5, …, 16, 32).

4.3. Lesion Frequency Analysis for Feature Voxel Selection

Feature selection is the most important process for SVM. A chi-square test was performed for individual lesion maps between RRMS and NMOSD groups in the training set to evaluate the importance of discrimination power at a voxel. The voxels were included as features for SVM in order of significance level from the chi-square test.

Various number of voxels were tested to evaluate the optimal number of feature voxels for RRMS and NMOSD classification (from one voxel to whole brain voxels in order of significance level from the chi-square test in the training set; n = 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, and 48193 voxels).

5. Statistical Analysis

5.1. Data Division

FLAIR images were randomly divided into 70% of training set to build a machine model and 30% of test set to evaluate the performance of decisions from the human raters and SVM for each RRMS and NMOSD group, respectively.

5.2. Performance Measurements

Performance was measured by sensitivity (i.e., correct ratio for RRMS group), specificity (correct ratio for NMOSD group), accuracy (i.e., correct ratio regardless of group), and area under the curve (AUC) from the decisions on test set by the human raters (i.e., rater A and B) and various SVM models, depending on number of feature voxels.

5.3. Intra-Rater Reliability of Human Raters

Cohen's kappa (κ) test was adopted to evaluate the intra-rater reliability of human raters between first and second decision on randomly duplicated 10% of the same sample in test set.

5.4. Inter-Rater Agreement Between Human Raters and SVM

McNemar's and Cohen's kappa (κ) tests were used to test the inter-rater agreement between decisions on test samples from the human raters A and B, SVM, and each other. Strength of agreement based on κ was judged according to the following guidelines: < 0.2 = slight; 0.2–0.4 = fair; 0.4–0.6 = moderate; 0.6–0.8 = substantial; > 0.8 = almost perfect.[30)]

Results

1. Baseline characteristics of patients with RRMS and NMOSD with AQP4-IgG

We reviewed 748 axial FLAIR MRIs from 172 patients with RRMS and 293 axial FLAIR MRIs from 97 patients with NMOSD with AQP4-IgG. We excluded two MRIs in patients with RRMS and one MRI in patients with NMOSD with AQP4-IgG due to poor image quality. Thus, the final analysis included 746 and 292 MRIs from 172 patients with RRMS and 97 patients with NMOSD with aquaporin-4 immunoglobulin G, respectively.

The clinical characteristics of all patients with RRMS and NMOSD with AQP4-IgG are presented in Table 1. The mean age at onset was lower in the patients with RRMS (32.7 ± 12.4 years) than in those with NMOSD (39.3 ± 12.9 years). Forty-four (25.6 %) of the RRMS patients and 12 (12.4%) of the NMOSD patients were male. The disease duration in both groups was comparable (RRMS = 10.6 ± 6.5 years, NMOSD = 9.7 ± 6.0 years).

Table 1. Clinical characteristics of patients with RRMS and NMOSD with AQP4-IgG

	RRMS (n = 172)	NMOSD (n = 97)	*P*-value
Age at onset, mean ± SD	32.7 ± 12.4	39.3 ± 12.9	< 0.001
Male, n (%)	44 (25.6)	12 (12.4)	0.010
Disease duration, mean ± SD	10.6 ± 6.5	9.7 ± 6.0	0.270
AQP4 IgG, n (%)	0 (0)	97 (100)	< 0.001
Axial FLAIR MRI, n	746	292	-

RRMS: relapsing-remitting multiple sclerosis, NMOSD: neuromyelitis optica spectrum disorder, AQP4-IgG: aquaporin-4 immunoglobulin G, FLAIR MRI: fluid-attenuated inversion recovery magnetic resonance image, SD: standard deviation.

2. Lesion frequency analysis of brain MRI from RRMS and NMOSD patients

The lesion frequency maps for the RRMS and NMOSD patient groups are shown in Figure 1A and 1B. The maps may include some artifacts like right cerebellum, which show high frequency in both diseases. Figure 1C shows the voxel-wise chi-squared test comparison of the lesion distribution in each group. Frequent lesions for RRMS were adjacent to the lateral ventricle and in the inferior temporal lobe. In contrast, frequent lesions for NMOSD were in the dorsal medulla, cerebral peduncle/internal capsule, corpus callosum, and nonspecific subcortex.

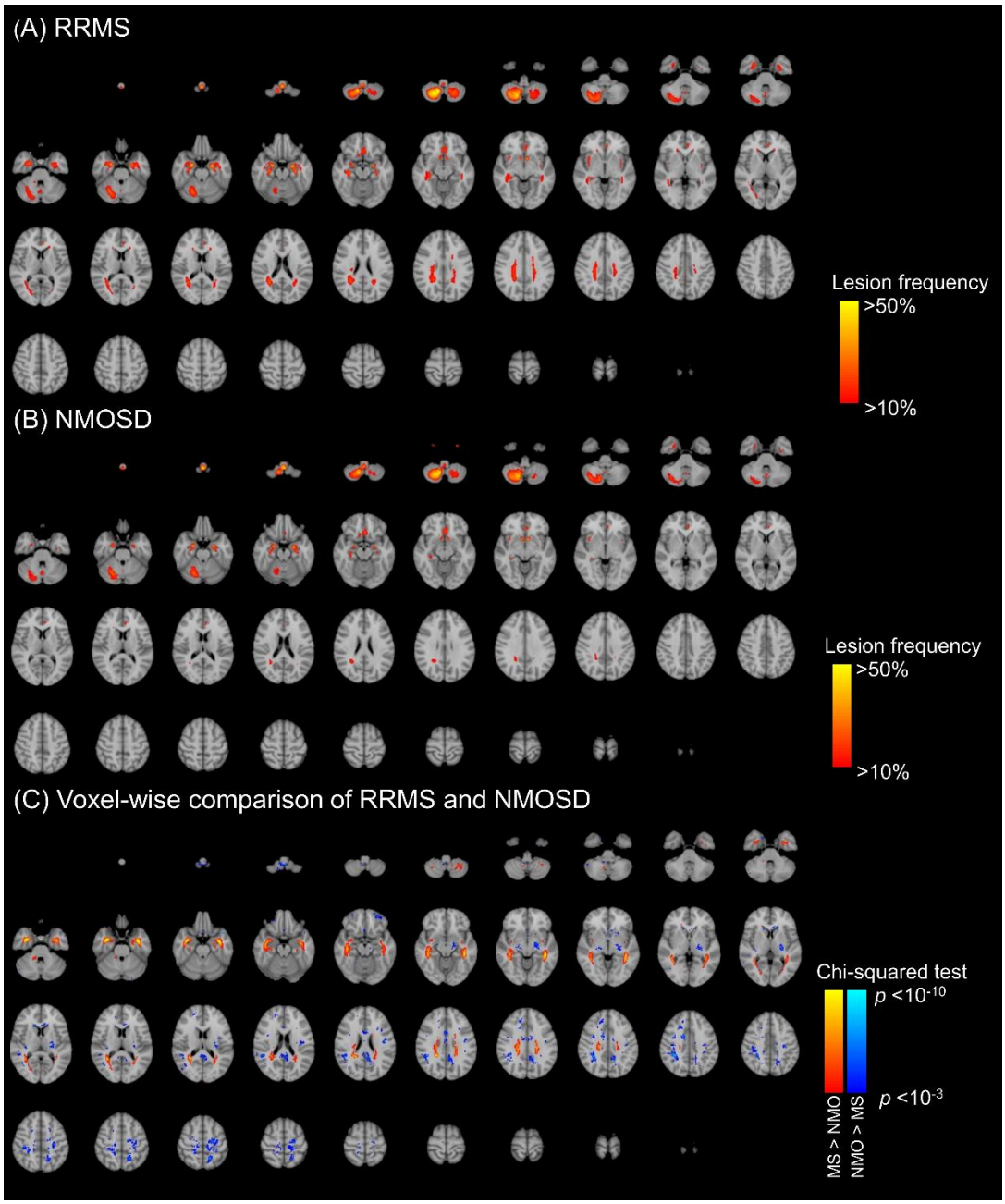


Figure 1. Lesion frequency maps for the relapsing-remitting multiple sclerosis (RRMS) and neuromyelitis optica spectrum disorder (NMOSD)

Lesion frequency maps for 172 patients with RRMS (A) and 97 patients with NMOSD (B). The color scale (from > 10% to > 50%) represents the lesion frequency in a spatial location. (C) Voxel-wise comparison of lesion frequency map with chi-squared test. The red color scale represents the frequent sites for RRMS, and the blue color scale represents the frequent sites for NMOSD.

3. Diagnostic performance of SVM

We randomly divided brain MRIs to the training (70%) and the test (30%) sets with the same percentage of patients with RRMS and NMOSD. The baseline characteristics of the training and test sets were comparable (Supplementary Table 1). Figure 2 shows schematic workflow of machine learning. The voxels were selected as features for SVM in order of significance level from the chi-square test. Various number of voxels were tested from one voxel to whole brain voxels in order of significance level from the chi-square test in the training set; n = 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, and 48193 voxels. Figure 3 presents the representative images of selected voxel distribution (Figure 3A = 2 voxel, Figure 3B = 32 voxel and, Figure 3C = 4096 voxel). Figure 4 summarizes the performance of the SVM. To achieve the best AUC of the test set, the SVM model from 2^7 voxel was selected. The performance of SVM from 2^7 voxel was 57.5% sensitivity, 78.4% specificity, 63.2% accuracy, and 75.3% AUC. Figure 5B shows the distribution of selected 2^7 voxels. The selected voxel lesions were dorsal medulla, inferior temporal lobe, adjacent to lateral ventricle, and nonspecific subcortex, which largely agreed with the lesions from the chi-squared testing comparison (Figure 5A).

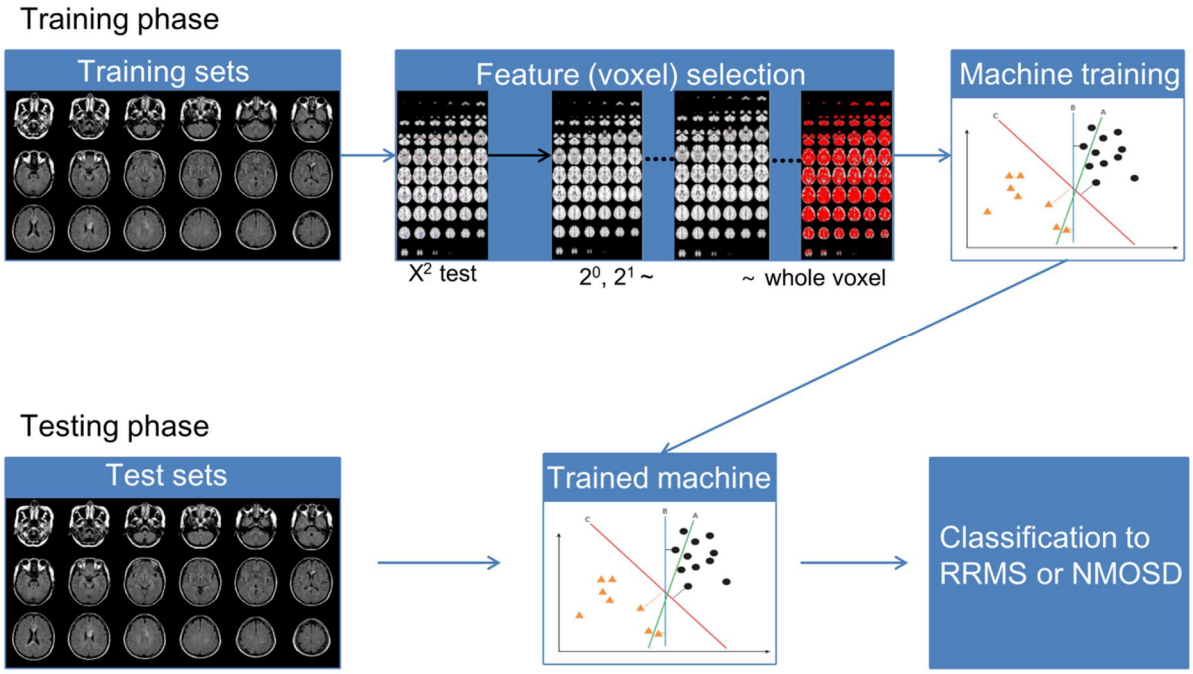


Figure 2. Schematic workflow of the machine learning

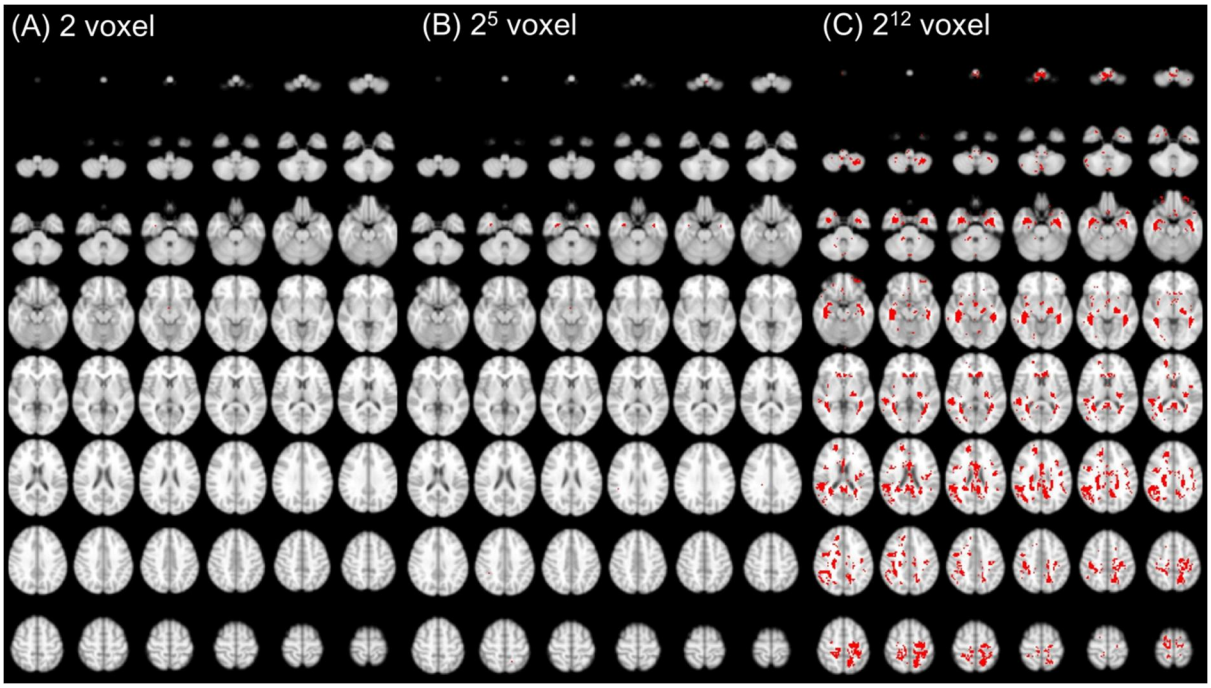


Figure 3. Representative images of selected voxel distribution

The voxels were selected as features for SVM in order of significance level from the chi-square test. Various number of voxels were tested from one voxel to whole brain voxels in order of significance level from the chi-square test in the training set; n = 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, and 48193 voxels. These are representative images of selected voxel distribution: (A) 2 voxel, (B) 32 voxel, and (C) 4096 voxel.

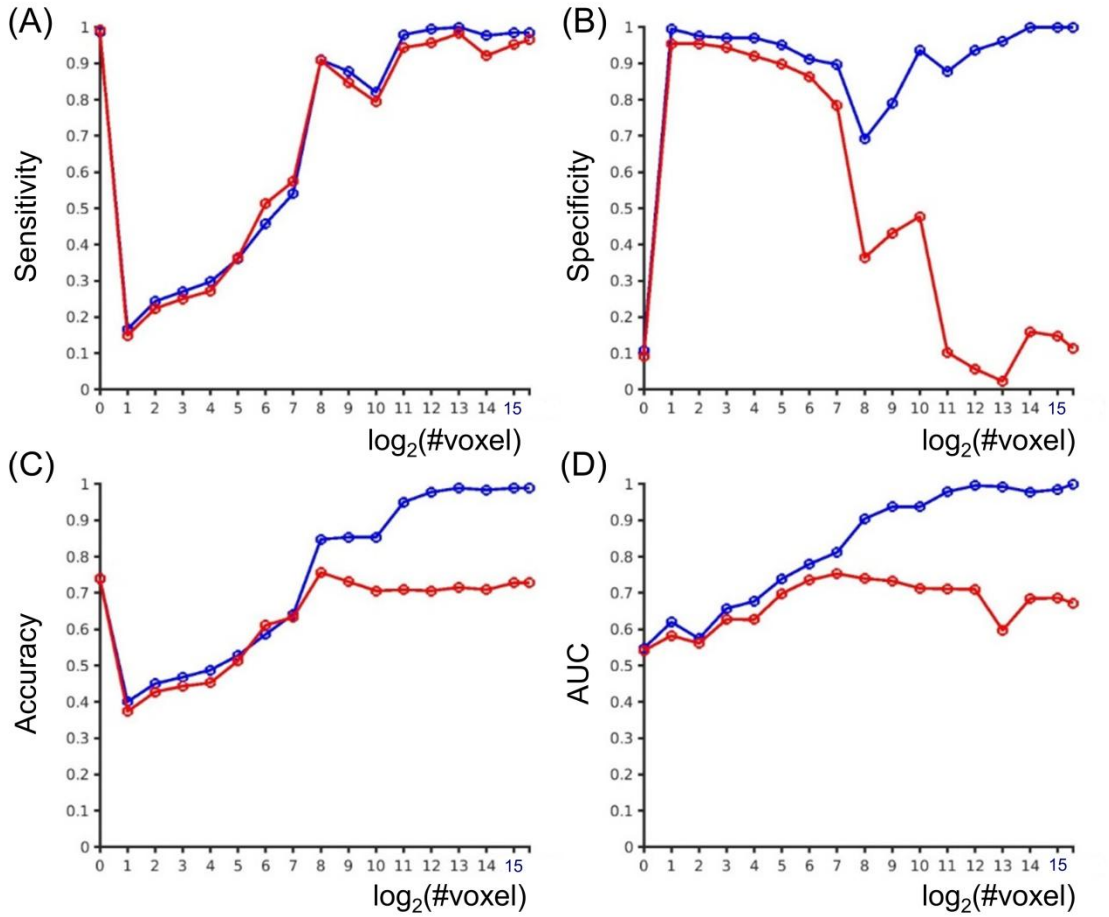


Figure 4. Diagnostic performance of support vector machine (SVM)

Blue lines show the performance with the training set. Red lines show the performance with the test set. The relationship between the number of voxels selected for SVM and sensitivity (A), specificity (B), accuracy (C), and area under curve (AUC) (D) are depicted, respectively. The highest AUC was achieved with the 2^7 voxel-selected SVM model.

#voxel: number of voxels, AUC: area under curve

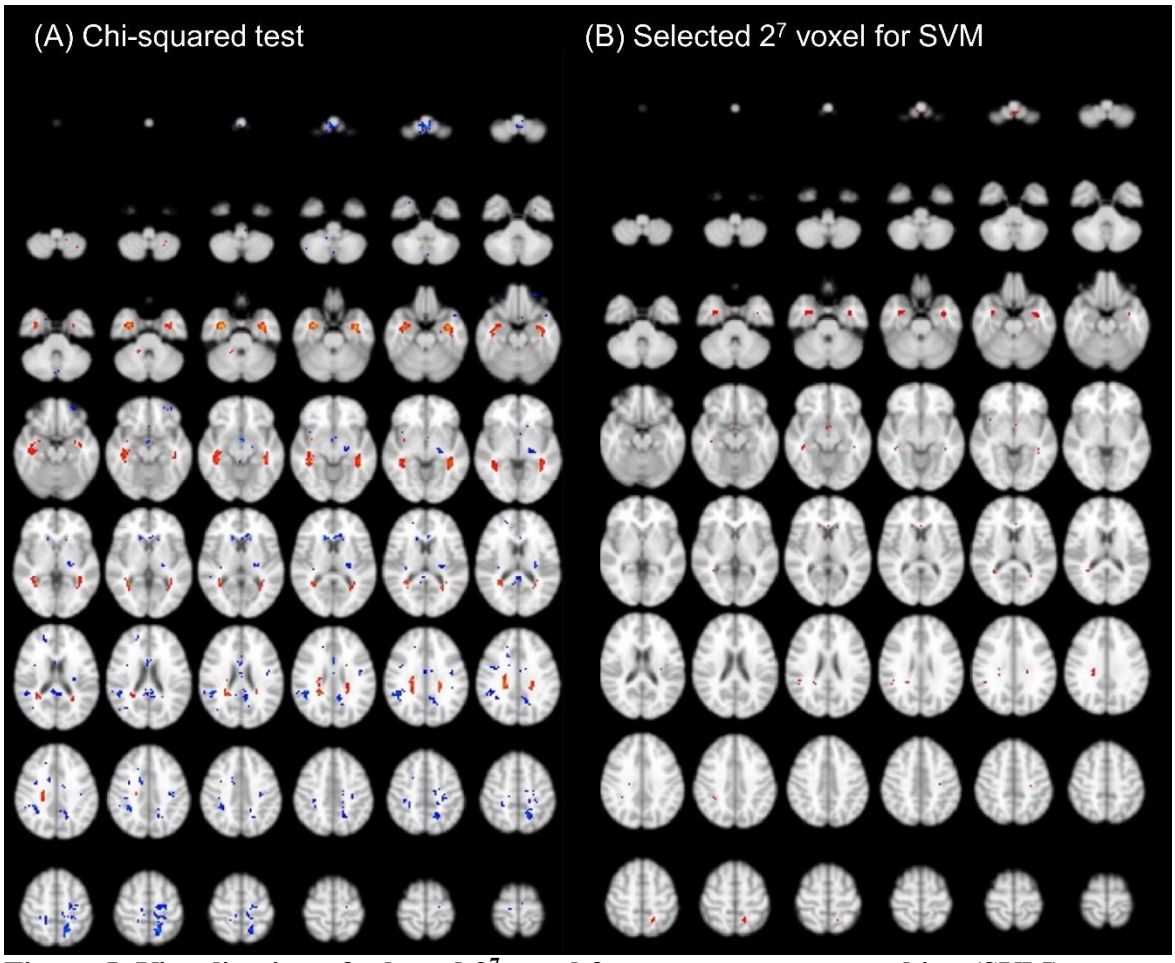


Figure 5. Visualization of selected 2^7 voxel for support vector machine (SVM) compared to chi-squared testing comparison

(A) Voxel-wise comparison of lesion frequency map with chi-squared test. (B) Visualization of selected 2^7 voxel for support vector machine (SVM).

4. Diagnostic performance of human raters and SVM

The performance of human raters and SVM is shown in Table 2. The specificity of SVM was comparable to or better than that of human raters. The sensitivity of SVM was lower than that of human raters. Intra-rater reliability of each human rater was substantial (κ, rater A = 0.680, rater B = 0.679); inter-rater agreement between the human raters was moderate (κ = 0.414); and inter-rater agreement between the SVM and each human rater was fair (κ = 0.279 for SVM and rater A, κ = 0.262 for SVM and rater B). McNemar's test suggested no systematic difference of accuracy between SVM and each human rater (p = 0.111 for SVM and rater A, p = 0.407 for SVM and rater B).

Table 2. Diagnostic performance of support vector machine (SVM) and human raters

	Sensitivity	Specificity	Accuracy	Intrarater reliability
SVM	57.5 (51.0–63.7)	78.4 (68.7–85.7)	63.3 (57.9–68.4)	100.0
Rater A	64.0 (57.6–70.0)	81.8 (72.5–88.5)	69.0 (63.7–73.8)	68.0
Rater B	69.7 (63.5–75.3)	58.0 (47.5–67.7)	66.5 (61.1–71.4)	67.9

Data are presented as percentage (95% confidence interval).

Discussion

1. Diagnostic criteria of MS and NMOSD

Diagnostic criteria for MS have evolved over time, with the most recent being the 2017 McDonald criteria from the International Panel on Diagnosis of Multiple Sclerosis (1965 Schumacher, 1983 Poser, 2001, 2005, 2010 McDonald).[31-36] The increasing incorporation of paraclinical assessments, especially imaging, to supplement clinical findings has allowed earlier, more sensitive, and more specific diagnosis.[37] The 2017 McDonald criteria were intended to simplify or clarify components of the 2010 McDonald criteria, to facilitate earlier diagnosis when multiple sclerosis was likely but not diagnosable with the 2010 McDonald criteria, and to preserve the specificity of the 2010 McDonald criteria and promote their appropriate application to reduce the frequency of misdiagnosis.[36] A recent study applying the 2017 McDonald criteria to a patient with typical clinically-isolated syndrome (CIS) showed greater sensitivity (68% vs. 36%) but less specificity (61% vs. 85%) for a second occurrence of symptoms than the 2010 McDonald criteria.[38] Another study that applied the new criteria to Korean patients with CIS showed higher sensitivity (89% vs. 53%) but lower specificity (43% vs. 69%) compared with the 2010 McDonald criteria for prediction of conversion to clinically-definite MS.[39] Therefore, the previous diagnostic criteria for MS were intended for earlier diagnosis of patients in whom MS was clinically suspected, rather than to differentiate MS from other mimicking disorders.[37] With the current criteria, clinicians should differentiate various other CNS diseases when they encounter a patient with abnormal brain MRI, considering clinical and laboratory information in addition to MRI characteristics.

The previously reported brain lesion distribution criteria for distinguishing MS from seropositive NMOSD showed high sensitivity (92%) and specificity (96%).[17] The suggested criteria was: "at least one lesion adjacent to the body of the lateral ventricle and in the inferior temporal lobe; or the presence of a subcortical U-fiber lesion; or a Dawson's finger-type lesion," which could be somewhat subjective and difficult to apply for clinicians who had minimal experience with radiological findings from patients with MS or NMOSD.

2. Advantages of the study

In this study, we demonstrated that machine learning could identify RRMS and NMOSD with acceptable accuracy comparable to neurologists with more than two years of clinical experience with demyelinating disease. The machine-learning method for differential diagnosis offers some advantages, including consistent interpretation (intra-rater reliability $\kappa = 1.0$), moderately high sensitivity and specificity, and prompt reporting of results.

Matthews et al. reported brain lesion distribution criteria but only analyzed MRIs with brain abnormalities.[17] Jurynczyk et al. applied the same criteria to distinguish MS from AQP4-Ab NMOSD and MOG-antibody disease, but they only included MRIs with the presence of brain lesions.[20] In the present study, we included all brain MRIs with or without abnormal lesions. A total of 94% (44/47) of the brain MRIs without abnormalities from the NMOSD test set were classified as NMOSD, suggesting the SVM algorithm identified the brain MRIs without abnormalities as NMOSD. Therefore, we confirmed that machine learning could overcome the limitation of the previous brain lesion distribution criteria, which excluded brain MRIs without abnormalities.

In Europe and North America, prevalence of MS (100 per 100,000) outweighs that of NMOSD (5 or less per 100,000).[4] In contrast, in Asia where prevalence of MS (0–20 per

100,000) is relatively low [40] and relative frequency of NMO to MS is 1.06,[41] differential diagnosis of the two diseases is particularly important. Moreover, because several disease-modifying treatments for MS can aggravate the course of NMOSD,[14-16] high diagnostic accuracy for NMOSD in our SVM model is beneficial for clinical practice in Asian countries. Our machine-learning method can discern RRMS and NMOSD with only brain MRI data, in contrast to the human raters with comparable diagnostic performance even with additional clinical information (age at onset, age at MRI, sex, and disease duration).

Therefore, machine learning may serve as a diagnostic support system for a clinician who has minimal experience with RRMS and NMOSD.

3. Limitation of the study

Machine learning (i.e., SVM) for discriminating medical diagnoses has been previously reported. Salvatore et al. used an SVM algorithm on T1-weighted brain MRIs in Parkinson's disease and progressive supranuclear palsy, with more than 90% accuracy, specificity, and sensitivity of diagnosis.[21] Magnin et al. applied an SVM algorithm to classify patients with Alzheimer's disease and control subjects with 96.6% mean specificity and 91.5% mean sensitivity.[25] Considering the strong performance (more than 90%) of the previous studies, our result of 57.5% sensitivity and 78.4% specificity seems relatively low. First, the low diagnostic performance may reflect a genuine difficulty in discriminating between RRMS and NMOSD. In our study, the diagnostic performance of board-certified neurologists who had more than two years of clinical experience with demyelinating disease showed similar sensitivity and specificity compared to machine learning. Second, through reviewing the false-negative images of our SVM model, we recognized that automatic lesion identification could be the main cause of low sensitivity for RRMS. The automatic lesion identification

labeled many artifacts as true lesions, mainly those located at the periventricular regions, including adjacent areas to third and fourth ventricle and corpus callosum.[42, 43] These artifact-prone areas overlapped with the known susceptible sites for NMOSD. Therefore, in our SVM model, many brain MRIs from patients with RRMS that mislabeled artifacts as lesions were classified as NMOSD, even though there were small but obviously characteristic RRMS lesions (Supplementary Figure 1).

We can consider some changes in methodology to improve diagnostic performance of machine learning. First, we may change the image processing method. In this study, we defined the factitious hyperintensity threshold as more than 0.32 ($p < 0.01$) for lesion identification, which caused false inclusion of FLAIR artifacts. We may apply voxel hyperintensity as continuous variables for SVM training. Second, we may include clinical information as additional features for the machine-learning training.[44] Third, we may analyze only acute-phase MRIs (< 3 months after relapse). Fourth, we may apply a "deep-learning" method for this classification. Currently, deep-learning techniques are state-of-the-art for classification of images. To date, promising results have been reported, including detection of pulmonary tuberculosis from chest radiography, diabetic retinopathy from fundus images, and skin cancer from clinical images.[45-47] We may apply the deep convolutional neural network for classification of MRIs from patients with RRMS and NMOSD, comparing the diagnostic performance to that from the SVM model in this study.

Finally, the current machine-learning algorithm can only be applied for dichotomous differentiation between RRMS and NMOSD after excluding other CNS diseases. By using clinical and laboratory data, clinicians need to exclude other CNS disorders that mimic demyelinating disease before applying the machine-learning algorithm.

4. Perspective of the machine learning for medical images

Today's machine-learning approaches have been progressing very rapidly. Because of the rapid pace of technological advancements, tasks previously thought to be limited to clinicians will be acquired by machine-learning systems. Machine learning is already being applied in the practice of radiology, and these applications will probably grow at a rapid pace [23]. There are several neurologic diseases that can be aided by machine-learning systems for diagnosis: primary CNS lymphoma,[48] primary angiitis of the CNS,[49] and other inflammatory diseases. These diseases need confirmatory biopsy, but the machine-learning system may obviate the need for invasive biopsy after being trained by an adequate amount of imaging data.

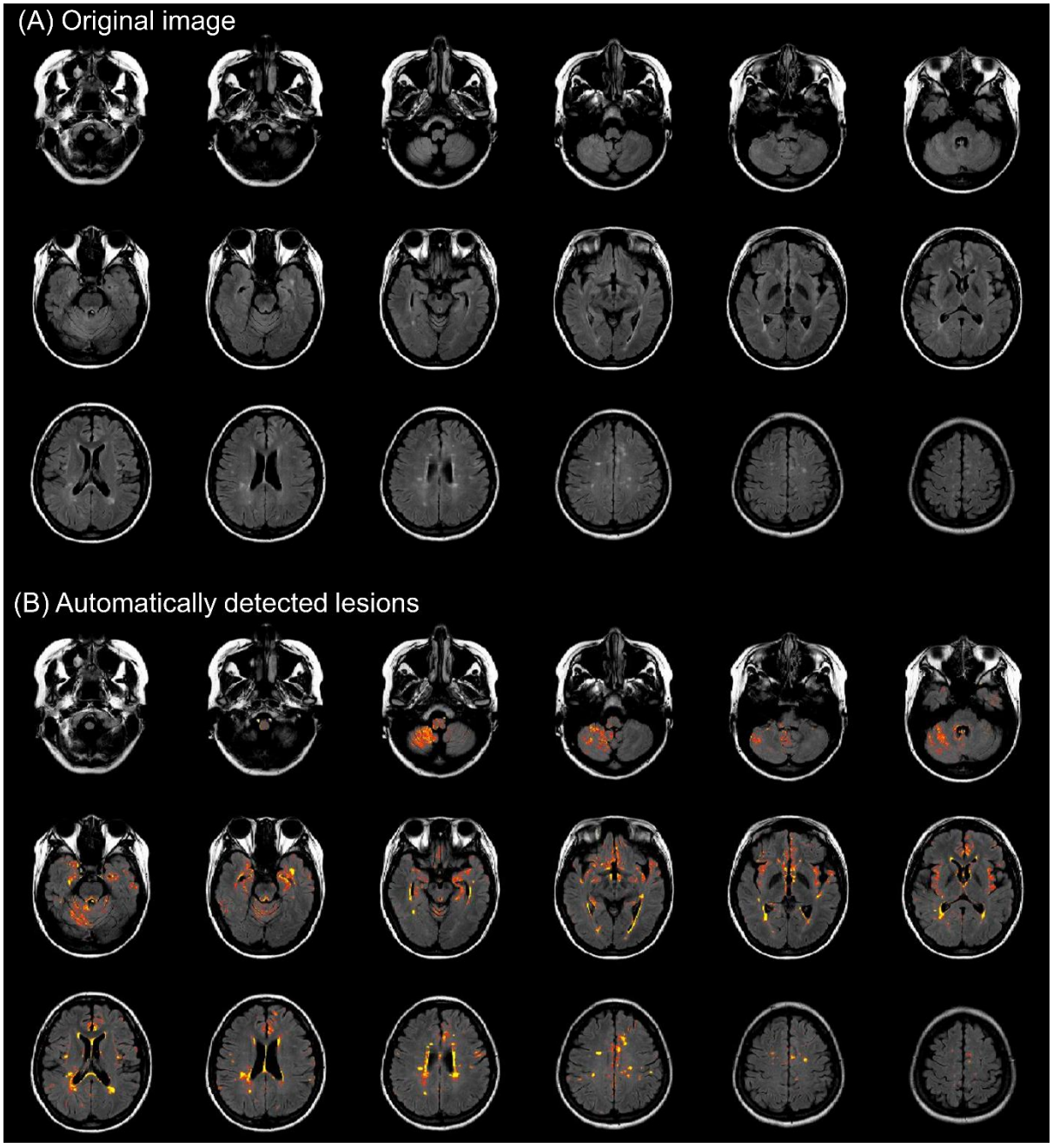
5. Conclusion

In conclusion, we confirmed that the machine learning method on brain MRI data can discern RRMS and NMOSD with comparable accuracy to that of clinicians. The machine learning technique may aid differential diagnosing the two important demyelinating diseases in clinical practice.

Supplementary Table 1. Baseline characteristics of training and test sets

	Training set (n=722)	Test set (n=316)	*P*-value
Attack index, mean ± SD	3.2 ± 2.4	3.2 ± 2.8	0.869
Disease duration, mean ± SD	5.3 ± 5.9	5.2 ± 5.5	0.714
Age at onset, mean ± SD	34.6 ± 12.1	33.7 ± 12.1	0.301
Age at imaging, mean ± SD	38.4 ± 11.7	37.2 ± 11.9	0.142
Male, n (%)	121 (16.8)	60 (19.0)	0.384
Patients with RRMS, n (%)	518 (71.7)	228 (72.2)	0.893
Patients with NMOSD, n (%)	204 (28.3)	88 (27.8)	

RRMS: relapsing-remitting multiple sclerosis, NMOSD: neuromyelitis optica spectrum disorder, SD: standard deviation.



Supplementary Figure 1. Representative false-negative MRI (RRMS, but classified as NMOSD by the SVM model)

(A) Original image, (B) Automatically detected lesions. It captured many artifacts as lesions; for example, cerebellum, medulla, and adjacent area to third ventricle, in addition to Dawson finger-like true lesions.

RRMS: relapsing-remitting multiple sclerosis, NMOSD: neuromyelitis optica spectrum disorder, SVM: support vector machine..

Reference

1. Wingerchuk DM, Banwell B, Bennett JL, Cabre P, Carroll W, Chitnis T, et al. International consensus diagnostic criteria for neuromyelitis optica spectrum disorders. Neurology 2015;85(2):177-89.
2. Flanagan EP, Cabre P, Weinshenker BG, Sauver JS, Jacobson DJ, Majed M, et al. Epidemiology of aquaporin-4 autoimmunity and neuromyelitis optica spectrum. Ann Neurol 2016;79(5):775-83.
3. Houzen H, Kondo K, Niino M, Horiuchi K, Takahashi T, Nakashima I, et al. Prevalence and clinical features of neuromyelitis optica spectrum disorders in northern Japan. Neurology 2017;89(19):1995-2001.
4. Mori M, Kuwabara S, Paul F. Worldwide prevalence of neuromyelitis optica spectrum disorders. J Neurol Neurosurg Psychiatry 2018;89(6):555-6.
5. Pandit L, Asgari N, Apiwattanakul M, Palace J, Paul F, Leite MI, et al. Demographic and clinical features of neuromyelitis optica: A review. Mult Scler 2015;21(7):845-53.
6. Lennon VA, Wingerchuk DM, Kryzer TJ, Pittock SJ, Lucchinetti CF, Fujihara K, et al. A serum autoantibody marker of neuromyelitis optica: distinction from multiple sclerosis. Lancet 2004;364(9451):2106-12.
7. Wingerchuk DM, Lennon VA, Lucchinetti CF, Pittock SJ, Weinshenker BG. The spectrum of neuromyelitis optica. Lancet Neurol 2007;6(9):805-15.
8. Chan KH, Tse CT, Chung CP, Lee RL, Kwan JS, Ho PW, et al. Brain involvement in neuromyelitis optica spectrum disorders. Arch Neurol 2011;68(11):1432-9.
9. Kim W, Park MS, Lee SH, Kim SH, Jung IJ, Takahashi T, et al. Characteristic brain magnetic resonance imaging abnormalities in central nervous system aquaporin-4

autoimmunity. Mult Scler 2010;16(10):1229-36.

10. Trebst C, Raab P, Voss EV, Rommer P, Abu-Mugheisib M, Zettl UK, et al. Longitudinal extensive transverse myelitis--it's not all neuromyelitis optica. Nat Rev Neurol 2011;7(12):688-98.

11. Wingerchuk DM, Hogancamp WF, O'Brien PC, Weinshenker BG. The clinical course of neuromyelitis optica (Devic's syndrome). Neurology 1999;53(5):1107-14.

12. Wingerchuk DM, Lennon VA, Pittock SJ, Lucchinetti CF, Weinshenker BG. Revised diagnostic criteria for neuromyelitis optica. Neurology 2006;66(10):1485-9.

13. Kim S-M, Kim S-J, Lee HJ, Kuroda H, Palace J, Fujihara K. Differential diagnosis of neuromyelitis optica spectrum disorders. Therapeutic advances in neurological disorders 2017;10(7):265-89.

14. Kleiter I, Hellwig K, Berthele A, Kumpfel T, Linker RA, Hartung HP, et al. Failure of natalizumab to prevent relapses in neuromyelitis optica. Arch Neurol 2012;69(2):239-45.

15. Min JH, Kim BJ, Lee KH. Development of extensive brain lesions following fingolimod (FTY720) treatment in a patient with neuromyelitis optica spectrum disorder. Mult Scler 2012;18(1):113-5.

16. Palace J, Leite MI, Nairne A, Vincent A. Interferon Beta treatment in neuromyelitis optica: increase in relapses and aquaporin 4 antibody titers. Arch Neurol 2010;67(8):1016-7.

17. Matthews L, Marasco R, Jenkinson M, Kuker W, Luppe S, Leite MI, et al. Distinction of seropositive NMO spectrum disorder and MS brain lesion distribution. Neurology 2013;80(14):1330-7.

18. Kim HJ, Paul F, Lana-Peixoto MA, Tenembaum S, Asgari N, Palace J, et al. MRI

characteristics of neuromyelitis optica spectrum disorder: an international update. Neurology 2015;84(11):1165-73.

19. Huh SY, Min JH, Kim W, Kim SH, Kim HJ, Kim BJ, et al. The usefulness of brain MRI at onset in the differentiation of multiple sclerosis and seropositive neuromyelitis optica spectrum disorders. Mult Scler 2014;20(6):695-704.

20. Jurynczyk M, Tackley G, Kong Y, Geraldes R, Matthews L, Woodhall M, et al. Brain lesion distribution criteria distinguish MS from AQP4-antibody NMOSD and MOG-antibody disease. J Neurol Neurosurg Psychiatry 2017;88(2):132-6.

21. Salvatore C, Cerasa A, Castiglioni I, Gallivanone F, Augimeri A, Lopez M, et al. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. J Neurosci Methods 2014;222:230-7.

22. Giger ML. Machine Learning in Medical Imaging. J Am Coll Radiol 2018;15(3 Pt B):512-20.

23. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. Radiographics 2017;37(2):505-15.

24. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the Brain: Artificial Intelligence in Stroke Imaging. J Stroke 2017;19(3):277-85.

25. Magnin B, Mesrob L, Kinkingnehun S, Pelegrini-Issac M, Colliot O, Sarazin M, et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 2009;51(2):73-83.

26. Adams JG, Melhem E. Clinical usefulness of T2-weighted fluid-attenuated inversion recovery MR imaging of the CNS. AJR American journal of roentgenology 1999;172(2):529-36.

27. Gawne-Cain M, O'riordan J, Thompson A, Moseley I, Miller D. Multiple sclerosis

lesion detection in the brain: a comparison of fast fluid-attenuated inversion recovery and conventional T2-weighted dual spin echo. Neurology 1997;49(2):364-70.

28. Rovira A, Wattjes MP, Tintore M, Tur C, Yousry TA, Sormani MP, et al. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-clinical implementation in the diagnostic process. Nat Rev Neurol 2015;11(8):471-82.

29. Schmidt P, Gaser C, Arsic M, Buck D, Forschler A, Berthele A, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. Neuroimage 2012;59(4):3774-83.

30. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology 2010;73(9):1167-79.

31. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 2001;50(1):121-7.

32. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol 2011;69(2):292-302.

33. Polman CH, Reingold SC, Edan G, Filippi M, Hartung HP, Kappos L, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". Ann Neurol 2005;58(6):840-6.

34. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC, et al. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983;13(3):227-31.

35. Schumacher GA, Beebe G, Kibler RF, Kurland LT, Kurtzke JF, McDowell F, et al. Problems of experimental trials of therapy in multiple sclerosis: report by the panel on the evaluation of experimental trials of therapy in multiple sclerosis. Ann N Y Acad Sci 1965;122:552-68.

36. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol 2018;17(2):162-73.

37. Brownlee WJ, Hardy TA, Fazekas F, Miller DH. Diagnosis of multiple sclerosis: progress and challenges. Lancet 2017;389(10076):1336-46.

38. van der Vuurst de Vries RM, Mescheriakova JY, Wong YYM, Runia TF, Jafari N, Samijn JP, et al. Application of the 2017 Revised McDonald Criteria for Multiple Sclerosis to Patients With a Typical Clinically Isolated Syndrome. JAMA Neurol 2018;75(11):1392-8.

39. Hyun JW, Kim W, Huh SY, Park MS, Ahn SW, Cho JY, et al. Application of the 2017 McDonald diagnostic criteria for multiple sclerosis in Korean patients with clinically isolated syndrome. Mult Scler 2018:1352458518790702.

40. Cheong WL, Mohan D, Warren N, Reidpath DD. Multiple Sclerosis in the Asia Pacific Region: A Systematic Review of a Neglected Neurological Disease. Front Neurol 2018;9:432.

41. Kim SM, Waters P, Woodhall M, Yang JW, Yang H, Kim JE, et al. Characterization of the spectrum of Korean inflammatory demyelinating diseases according to the diagnostic criteria and AQP4-Ab status. BMC Neurol 2014;14:93.

42. Neema M, Guss ZD, Stankiewicz JM, Arora A, Healy BC, Bakshi R. Normal findings on brain fluid-attenuated inversion recovery MR images at 3T. AJNR Am J

Neuroradiol 2009;30(5):911-6.

43. Lavdas E, Tsougos I, Kogia S, Gratsias G, Svolos P, Roka V, et al. T2 FLAIR artifacts at 3-T brain magnetic resonance imaging. Clin Imaging 2014;38(2):85-90.

44. Ion-Margineanu A, Kocevar G, Stamile C, Sima DM, Durand-Dubief F, Van Huffel S, et al. Machine Learning Approach for Classifying Multiple Sclerosis Courses by Combining Clinical Data with Lesion Loads and Magnetic Resonance Metabolic Features. Front Neurosci 2017;11:398.

45. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-8.

46. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama 2016;316(22):2402-10.

47. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017;284(2):574-82.

48. Alcaide-Leon P, Dufort P, Geraldo AF, Alshafai L, Maralani PJ, Spears J, et al. Differentiation of Enhancing Glioma and Primary Central Nervous System Lymphoma by Texture-Based Machine Learning. AJNR Am J Neuroradiol 2017;38(6):1145-50.

49. Beuker C, Schmidt A, Strunk D, Sporns PB, Wiendl H, Meuth SG, et al. Primary angiitis of the central nervous system: diagnosis and treatment. Ther Adv Neurol Disord 2018;11:1756286418785071.

국문 요약

서론: 시신경척수염범주질환과 다발경화증의 치료법이 다르고, 다발경화증에 대한 질병조절치료가 시신경척수염범주질환을 악화시킬 수 있기 때문에, 두 질환을 감별하는 것은 중요하다. 뇌 자기공명영상 검사는 두 질환을 감별하는 중요한 진단법 중에 하나로, 현재까지 두 질환을 식별할 수 있는 뇌 자기공명영상검사의 특징을 찾기 위한 상당한 노력을 기울였다. 기계학습은 의료영상을 감별하는 한 가지 방법으로 연구되어 왔다. 이 연구의 목적은 뇌 자기공명영상검사 자료를 이용해 다발경화증과 시신경척수염범주질환의 감별진단을 수행할 수 있는 기계학습 방법을 구현하는 것이다.

연구방법: 우리는 2005 년부터 2017 년 사이에 서울아산병원에 내원한 재발-이장성 다발경화증과 아쿠아포린 4-면역글로불린 G (Aquaporin 4-immunoglobulin G)를 가진 시신경척수염범주질환 환자의 FLAIR (fluid attenuated inversion recovery) 자기공명영상 검사를 획득하였다. FLAIR 자기공명영상을 기계학습에 이용하였고, 기계학습은 병변 빈도 분석을 통해 복셀 (voxel)을 선택하고, 서포트 벡터 머신 (support vector machine)을 분류 알고리즘으로 사용하였다. 기계학습의 진단성능은 2 년 이상의 탈수초성질환 임상경험이 있는 2 명의 신경과 의사와 비교 되었다.

연구결과: 최종 분석에는 172 명의 재발-이장성 다발경화증 환자의 746 개 뇌 자기공명영상과 97 명의 시신경척수염범주질환 환자의 292 개 뇌 자기공명영상이 포함되었다. 병변 빈도 분석을 통해 가측 뇌실 근처와 하측두엽이 재발-이장성 다발경화증의 흔한 병변 위치이고, 등쪽 연수, 대뇌다리/속섬유막, 뇌들보가 시신경척수염범주질환의 흔한 병변 영역임을 확인하였다. 기계학습은 57.5% 민감도, 78.4% 특이도, 63.3% 정확도를 보였고, 이는 임상의사와 비교하였을 때, 상당한 수준의 일치도를 보였다. (Cohen's κ, rater A=0.279, and rater B=0.262).

결론: 결론적으로, 우리는 뇌 자기공명영상에 대한 기계학습이 임상 의사와 유사한 정확도로 재발-이장성 다발경화증과 시신경척수염범주질환을 감별 진단할 수 있다는 것을 확인하였다. 이는 임상 진료에서, 기계학습 보조 진단의 적용을 시사한다.

중심 단어: 뇌 자기공명영상, 다발경화증, 시신경척수염범주질환, 기계학습, 서포트 벡터 머신