Deep Learning-Based Method to Differentiate Neuromyelitis Optica Spectrum Disorder From Multiple Sclerosis

Hyunjin Kim [1†], Youngin Lee [2,3†], Yong-Hwan Kim [2], Young-Min Lim [1], Ji Sung Lee [4,5], Jincheol Woo [2], Su-Kyeong Jang [2], Yeo Jin Oh [1], Hye Weon Kim [1], Eun-Jae Lee [1], Dong-Wha Kang [1,2]* and Kwang-Kuk Kim [1]*

[1] Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea, [2] Asan Institute for Life Sciences, Asan Medical Center, Seoul, South Korea, [3] Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea, [4] Clinical Research Center, Asan Institute for Life Sciences, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea, [5] Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea

Background: Differentiating neuromyelitis optica spectrum disorder (NMOSD) from multiple sclerosis (MS) is crucial in the field of diagnostics because, despite their similarities, the treatments for these two diseases are substantially different, and disease-modifying treatments for MS can worsen NMOSD. As brain magnetic resonance imaging (MRI) is an important tool to distinguish the two diseases, extensive research has been conducted to identify the defining characteristics of MRI images corresponding to these two diseases. However, the application of such research in clinical practice is still limited. In this study, we investigate the applicability of a deep learning-based algorithm for differentiating NMOSD from MS.

Methods: In this study, we included 338 participants (213 patients with MS, 125 patients with NMOSD) who visited the Asan medical center between February 2009 and February 2020. A 3D convolutional neural network, which is a deep learning-based algorithm, was trained using fluid-attenuated inversion recovery images and clinical information of the participants. The performance of the final model in differentiating NMOSD from MS was evaluated and compared with that of two neurologists.

Results: The deep learning-based model exhibited an area under the receiver operating characteristic curve of 0.82 (95% CI, 0.75–0.89). It differentiated NMOSD from MS with an accuracy of 71.1% (sensitivity = 87.8%, specificity = 61.6%), which is comparable to that exhibited by the neurologists. The intra-rater reliability of the two neurologists was moderate (κ = 0.47, 0.50), which was in contrast with the consistent classification of the deep learning-based model.

Conclusion: The proposed model was verified to be capable of differentiating NMOSD from MS with accuracy comparable to that of neurologists, exhibiting the advantage of consistent classification. As a result, it can aid differential diagnosis between two important central nervous system inflammatory diseases in clinical practice.

INTRODUCTION

Neuromyelitis optica spectrum disorder (NMOSD) and multiple sclerosis (MS) are both inflammatory diseases of the central nervous system (CNS) (1). Because of the clinical and radiological similarities between the two, there has been a persistent debate on whether they are actually different. However, since the discovery of the anti-aquaporin-4 antibody (AQP4-Ab), which is an NMOSD-specific autoantibody (2), studies have confirmed MS and NMOSD to be distinct disease entities (3). Differentiating NMOSD from MS is of considerable importance in the field of diagnostics because the treatments for the two diseases differ considerably from each other, and disease-modifying therapies for MS can worsen NMOSD (4–6). Even though the presence of AQP4-Ab is essential for the diagnosis of AQP4-Ab-seropositive NMOSD, clinical and radiological differentiation between AQP4-Ab-seropositive NMOSD and MS remains crucial for the following reasons: (i) clinicians need to identify patients on whom the AQP4-Ab test should be performed; (ii) the result of an AQP4-Ab assay can be influenced by assay methodology and the patient's clinical status (3, 7, 8). Therefore, the role of brain MRI, which is the most common test for CNS inflammatory diseases, is significant in the differentiation between the two diseases (9).

Recently, machine learning-based algorithms have been applied to classify MRIs of patients with various neurological diseases (10–12). In particular, researchers have attempted to use various such methods, including multimodal data fusion and random forests, to differentiate NMOSD from MS (13, 14). Deep learning, which is a type of machine learning, does not require the specification of explicit features by experts and deduces most predictive features directly based on images (15, 16). Deep learning has been applied for classification based on medical images such as chest X-rays and fundus photographs, and it has been reported to exhibit excellent differentiation performance (17, 18). Fluid-attenuated inversion recovery (FLAIR) is an MRI technique that highlights T2 hyperintense lesions while suppressing cerebrospinal fluid (CSF) signals, thereby clearly revealing lesions that are in proximity to CSF, such as juxtacortical and periventricular lesions (19). FLAIR is considered to be superior to T2-weighted images in the context of the detection of MS brain lesions (20).

In this study, we aim to develop a deep learning model based on brain FLAIR MRIs and elementary clinical information to differentiate NMOSD from MS. Further, we evaluated the clinical applicability of the proposed model by comparing its performance with that of two neurologists.

MATERIALS AND METHODS

Participants

We retrospectively reviewed the medical records of patients with MS and NMOSD who visited the Asan Medical Center, Seoul, Korea, between February 2009 and February 2020. Patients with MS who fulfilled the 2010 McDonald criteria and those with NMOSD with AQP4 immunoglobulin G (AQP4-IgG) seropositivity, in accordance with the 2015 International Consensus of NMOSD, were included in this study. The AQP4-IgG-seropositive status was confirmed via a commercial fixed cell-based assay (Euroimmun, Lubeck, Germany). Patients without available brain MRI data were excluded from the study. As a representative MRI of each patient, the last MRI was examined to use the latest data. All images were deidentified prior to being transferred to the study investigators.
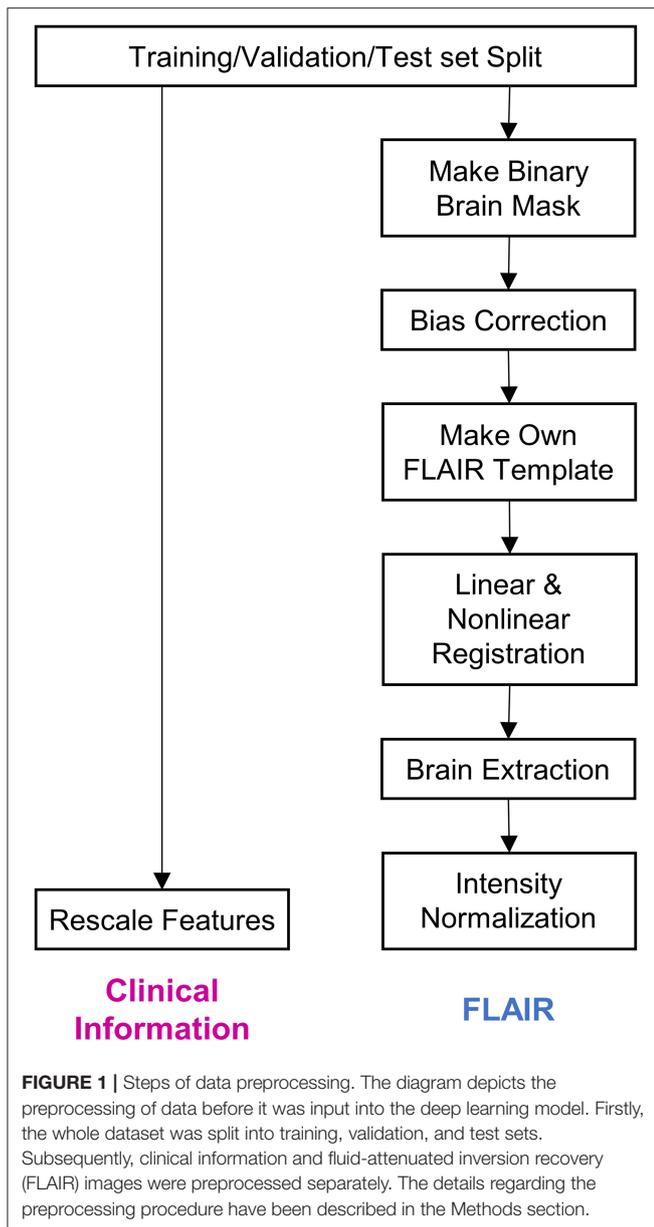
MRI Acquisition

2D FLAIR sequences were acquired using 1.5 T or 3.0 T scanners at the Asan Medical Center or other centers that referred patients and transferred images to the Asan Medical Center. The MRI protocol at the Asan Medical center is a fast spin-echo sequence with inversion recovery with the following scanning parameters: field of view $= 220 \times 200$ mm; voxel size $= 0.65$ mm $\times 0.82$ mm; a 336×231 acquisition matrix; time of repetition $= 11,000$ ms; time of echo $= 125$ ms; inversion time $= 2,500$ ms; slice thickness $= 5$ mm; slice gap $= 2$ mm; and number of acquisitions $= 2$.

NMOSD and MS Classification by Neurologists

Two board-certified neurologists (H.W. Kim and Y.J. Oh) who have completed a 1-year clinical fellowship in neuro-immunology participated in this study as the human raters. They binarily diagnosed MS and NMOSD independently of each other by reviewing the FLAIR images alongside the following clinical information of each patient: age at disease onset, age at the time of MRI, sex, disease duration, and duration from last relapse. We considered these five variables to be essential to adequately capture the elementary information required for differential diagnoses in clinical practice.

NMOSD and MS Classification via Deep Learning-Based Model

All methods were implemented using Python and PyTorch for deep learning. The algorithm was executed on the Intel Core i7-8700K 3.70 GHz processor with two Nvidia GeForce RTX 2080

FIGURE 1 | Steps of data preprocessing. The diagram depicts the preprocessing of data before it was input into the deep learning model. Firstly, the whole dataset was split into training, validation, and test sets. Subsequently, clinical information and fluid-attenuated inversion recovery (FLAIR) images were preprocessed separately. The details regarding the preprocessing procedure have been described in the Methods section.

to outliers, the features can be transformed to exhibit nearly identical scales while being minimally affected by outliers.

Image Pre-processing

Brain mask data for each FLAIR sequence were obtained using a previously reported brain extraction algorithm (21). The brain mask data are presented as binary data, wherein only the voxels corresponding to brain parts are coded to correspond to 1. The FLAIR images were bias-corrected via the lesion prediction algorithm (22), as implemented in lesion segmentation toolbox (LST) version 2.0.15 (www.statistical-modeling.de/lst.html), for statistical parametric mapping (SPM, www.fil.ion.ucl.ac.uk/spm/). Following bias-correction, we obtained a file containing deformation parameters for each image; the parameters were capable of registering the image from its native space to the Montreal Neurological Institute (MNI) space with a resolution of $2\,mm \times 2\,mm \times 2\,mm$ (23). Using these deformation parameters, we registered the MRIs belonging to the training set to the MNI space. Furthermore, using functional MRI of the brain software library (FSL) functions (24), these registered images were made to undergo 3D intensity normalization and were then averaged into one image, which was used as a template in this study. After creating the template, all the FLAIR images and brain masks were registered to the template and intensity normalization with Z-scoring was conducted.

Model Architecture

Deep learning, particularly convolutional neural networks (CNNs), has been widely used in computer vision (25). Among the different variants of CNNs, ResNet has exhibited remarkable performance in image classification (26). Additionally, although most imaging studies have used 2D CNNs as their model architecture, some recent studies have proposed the use of 3D CNNs to fully utilize the spatial features of MRI and achieve better performance (27–30). Therefore, we constructed a 3D CNN architecture based on the idea of ResNeXt (31), which is a more developed model based on ResNet (**Figure 2A**). To use both clinical information and FLAIR data simultaneously, we concatenated the clinical information at the end of CNN architecture and propagated the information through two fully connected layers (**Figure 2B**).

Modeling

We trained and validated several models that differ from each other in terms of learning rate scheduling strategies, dropout rates, loss functions, and data augmentation strategies. After validating them using the validation set, we selected the best model. Stochastic gradient descent with a momentum of 0.9 and a weight decay of 10^{-6} was used as the optimizer. The initial learning rate of 0.025 was decayed using cosine annealing, and the pre-defined minimal learning rate, 0.015, was reached after 20 epochs. This cycle was repeated throughout the whole training process. This training strategy is called stochastic gradient descent with warm restarts (32). A batch size of 16 was selected and focal loss was used as a loss function for the best model, in which gamma and alpha were 1 and 0.25, respectively. Focal loss is often used when the class ratio of data

Ti graphics processing units. **Figure 1** depicts the preprocessing steps whose details have been described below.

Subdivision of the Dataset Into Training, Validation, and Test Sets

We first split the dataset into training (45%, $n = 152$), validation (15%, $n = 51$), and test (40%, $n = 135$) datasets. We used stratified random sampling to ensure identical class ratios for every set.

Scaling Clinical Information

Four pieces of clinical information—age at disease onset, age at the time of MRI, disease duration, and duration from last relapse—were scaled to exhibit a median of 0 and an interquartile range of 1. As this scaling method uses statistics that are robust
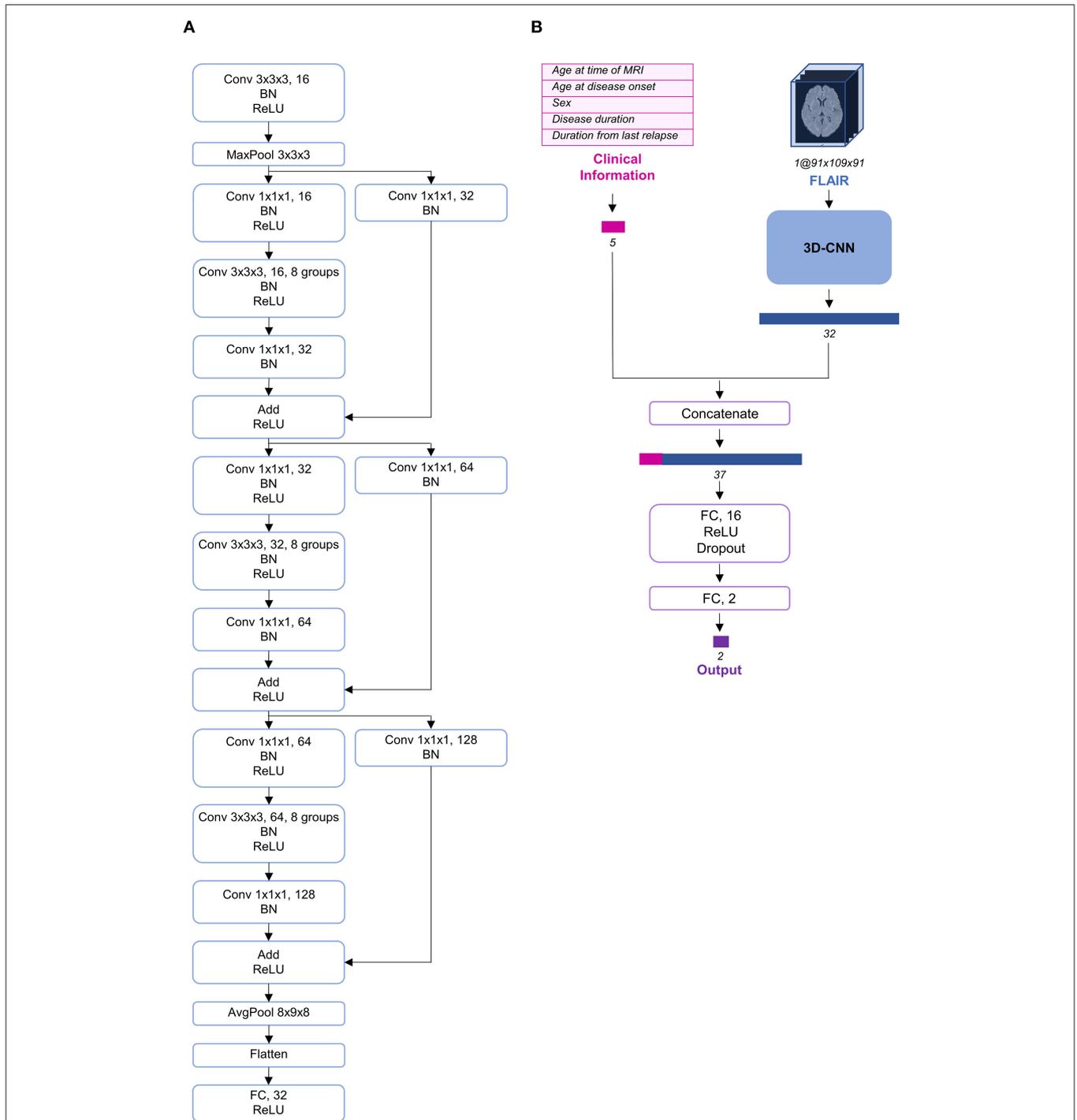
FIGURE 2 | Schema of the model architecture. **(A)** The 3D convolutional neural network architecture used in our study. **(B)** The entire structure of the proposed model. The preprocessed fluid-attenuated inversion recovery (FLAIR) image was transmitted through the 3D convolutional neural network (CNN) and its feature vector was extracted and concatenated with the preprocessed clinical information. AvgPool, average-pooling; BN, batch normalization; Conv, convolution; FC, fully connected; MaxPool, max-pooling; ReLU, rectified linear unit.

is imbalanced (33). We expected the model to focus more on the NMOSD class by training it using focal loss. To prevent overfitting, rotation or translation was randomly applied to

augment the FLAIR data immediately before they were input into the model. Further, the dropout strategy was used with a rate of 0.5. Both augmentation and dropout were conducted only
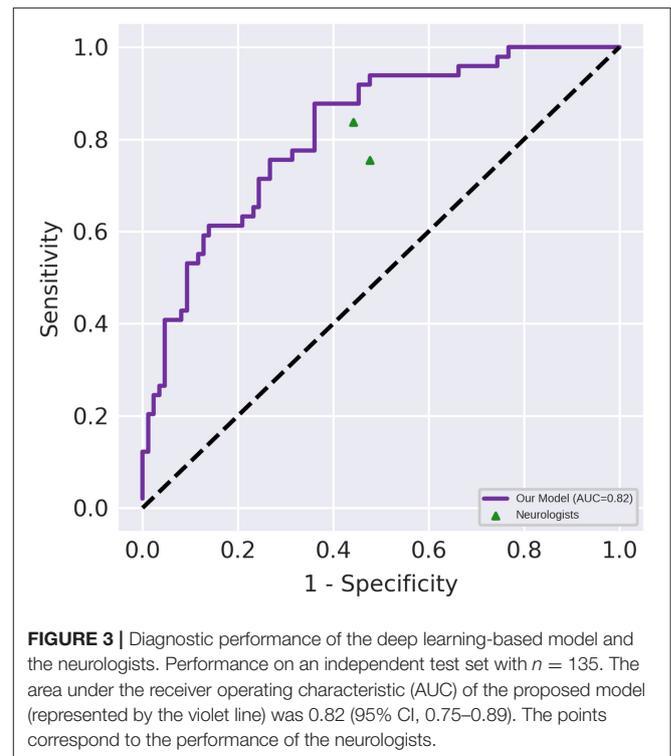
TABLE 1 | Baseline characteristics.

	MS	NMOSD	P-value
No.	213	125	
Male, n (%)	55 (25.8)	13 (10.4)	0.001
Age at onset, mean ± SD (years)	33.1 ± 12.3	41.7 ± 13.7	< 0.001
Age at imaging, mean ± SD (years)	37.1 ± 12.0	45.9 ± 13.2	<0.001
Disease duration, mean ± SD (years)	7.6 ± 6.6	5.3 ± 5.6	0.001
Duration from last relapse, mean ± SD (years)	3.6 ± 4.3	1.2 ± 2.1	<0.001
EDSS score, mean ± SD	2.4 ± 1.8	3.3 ± 1.8	<0.001
MRI performed at AMC, n (%)	192 (90.1)	105 (83.3)	0.095
MRI performed with 3 T scanner, n (%)	172 (80.8)	91 (72.2)	0.089

AMC, Asan medical center; EDSS, Expanded Disability Status Scale; NMOSD, neuromyelitis optica spectrum disorder; MS, multiple sclerosis; SD, standard deviation.

FIGURE 3 | Diagnostic performance of the deep learning-based model and the neurologists. Performance on an independent test set with n = 135. The area under the receiver operating characteristic (AUC) of the proposed model (represented by the violet line) was 0.82 (95% CI, 0.75–0.89). The points correspond to the performance of the neurologists.

during the training of the model and neither were applied during the validation or evaluation process. The final performance of the model was evaluated using the test set that had been split beforehand.

Statistical Analysis

In this study, descriptive summaries were represented as frequencies and percentages for categorical variables and as mean ± standard deviation for continuous variables. In order to compare pairs of groups, Student's t-tests were used for continuous variables and Chi-squared tests were used for categorical variables. The diagnostic performances of the proposed model and the human raters (i.e., raters A and B) were measured in terms of sensitivity (i.e., correct ratio for the NMOSD group), specificity (correct ratio for the MS group), and accuracy (i.e., correct ratio regardless of the group) of their decisions related to the test set. Regarding sensitivity, specificity, and accuracy, a two-sided 95% exact confidence interval (CI) was calculated using the Clopper–Pearson method (34). McNemar's test was used to compare the decision performances of the deep learning-based model and the human raters. The diagnostic performance of the deep learning-based model was assessed using area under the receiver operating characteristic curve (AUC). The 95% CI of AUC was calculated following DeLong's method (35). Cohen's kappa (κ) test was adopted to evaluate the intra-rater reliability of the human raters between the first and second decisions on randomly duplicated 10% of identical samples in the test set; moreover, the aforementioned test was also performed to test the inter-rater agreement (on the test set) between the decisions of the two human raters and between the decisions of the human raters and those of the deep learning-based model. The 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; and >0.8 = almost perfect (36). Two-sided P-values of <0.05 were indicative of statistical significance. All statistical analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC).

RESULTS

Baseline Characteristics of the Participants

We included 338 participants (213 patients with MS and 125 patients with NMOSD) in the study. The baseline characteristics of all participants have been presented in **Table 1**. The mean age at onset was lower for the MS group than that for the NMOSD group (33.1 ± 12.3 vs. 41.7 ± 13.7 years, $P < 0.001$). Fifty-five (25.8%) patients in the RRMS group and 13 (10.4%) in the NMOSD group were male. The mean disease duration and duration from last relapse were observed to be longer for the MS group. The percentages of MRIs performed at the Asan Medical Center (vs. other centers) and those of MRIs performed with a 3 T scanner (vs. 1.5 T scanner) were comparable between the two groups. Further, 30% (37/125) of brain MRIs from patients with NMOSD did not exhibit any T2 hyperintense lesions (≥3 mm). In contrast, all brain MRIs from patients with MS exhibited more than one T2 hyperintense lesion.

Diagnostic Performance

Figure 3 depicts the receiver operating characteristic curve of the model when it is evaluated using the test set (AUC, 0.82 [95% CI, 0.75–0.89]). The classification results of the human raters and deep learning-based model have been presented in **Table 2**. The accuracy of the deep learning model in differentiating NMOSD from MS was 71.1%, with a sensitivity of 87.8% and a specificity of 61.6%, all of which were comparable to those of the human raters. The intra-rater reliability of each human rater was moderate (κ of rater A = 0.471, κ of rater B = 0.500); the inter-rater agreement

TABLE 2 | Classification results.

	Accuracy, %	*P*-value	Sensitivity, %	*P*-value	Specificity, %	*P*-value
Deep learning-based model	71.1 (62.7–78.6)	-	87.8 (75.2–95.4)	-	61.6 (50.5–71.9)	-
Rater A	65.9 (57.3–73.9)	0.382	83.7 (70.3–92.7)	0.754	55.8 (44.7–66.5)	0.511
Rater B	60.7 (52.0–69.0)	0.081	75.5 (61.1–86.7)	0.180	52.3 (41.3–63.2)	0.280

Data have been presented with a 95% confidence interval. P-value for comparison with the deep learning-based model.

between the human raters was moderate ($\kappa = 0.437$); and the inter-rater agreements between the deep learning-based model and each human rater were fair and slight, respectively ($\kappa = 0.288$ between the deep learning-based model and rater A; $\kappa = 0.154$ between the deep learning-based model and rater B).

DISCUSSION

In this study, we demonstrated that the proposed deep learning-based model is capable of differentiating NMOSD from MS with an accuracy of 71.1% (sensitivity = 87.8%, specificity = 61.6%), which was comparable to that achieved by the neurologists. Although the diagnostic accuracies were comparable, the deep learning-based model offers some advantages over human evaluation, including greater consistency in classification (intra-rater reliability $\kappa = 1.0$) and prompt reporting of results. The intra-rater reliability of the human raters was only moderate ($\kappa = 0.47$–0.50).

The definition of a brain lesion distribution as "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" has been suggested as a criterion; this can be employed to distinguish patients with MS from those with NMOSD with 92% sensitivity and 96% specificity (37). In addition, the same criterion was used to distinguish MS from myelin oligodendrocyte glycoprotein antibody-associated diseases with a sensitivity of 90.9% and a specificity of 95.2% (38). However, the aforementioned studies did not include brain MRIs without T2 hyperintense lesions in their investigations, and the criterion included subjective definitions such as Dawson's fingers, which could be difficult to apply for clinicians who have little experience with CNS inflammatory diseases. In the current study, we included all MRIs with or without brain lesions and did not designate any specific lesion criteria. Instead, the deep learning-based model learned the most predictive features directly from the images. Therefore, we suggest that the proposed deep learning-based model is capable of overcoming the low practicability of the previously published brain lesion distribution criteria.

Two machine learning-based and one deep learning-based methods using brain MRI data had previously been developed to distinguish NMOSD from MS (13, 14, 39). The studies had adopted various methods, including multimodal data fusion, random forest classification, and hierarchical multimodal fusion, and achieved accuracies ranging from 74 to 88%. However, the applicability of the developed methods in clinical practice was uncertain owing to the lack of their comparison to

assessments by clinicians. In our study, we demonstrated that the proposed model exhibits a performance comparable to that of trained clinicians, demonstrating its potential clinical applicability. Recent systematic reviews of studies on the comparison between performances of artificial intelligence and clinicians have reported that the performance of artificial intelligence was comparable to that of clinicians (40, 41).

Several limitations should be noted for the current study. Firstly, we conducted this study in a retrospective manner without external validation, which entails the risk of bias and lack of generalizability. Further, the study participants were of a single ethnicity (Korean), which implies that our result might not be applicable to patients from other ethnic backgrounds (42). Future prospective studies that incorporate data from other international centers and larger data samples can overcome this limitation. Secondly, spinal MRI and CSF findings, other important diagnostic clues (7, 43), may strengthen the performance of the deep learning-based model. In the present study, we have tried to develop a model based on minimal information, but we may add these variables when developing a model in the future. Lastly, the proposed deep learning-based model was trained for binary classification. Hence, this model is not an automatic brain MRI interpreter. Our results provide evidence that deep learning can support the objective differential diagnosis of MS and NMOSD. In the future, we may attempt to include other CNS diseases using this deep learning-based model to solve multiclass classification problems.

In conclusion, the proposed deep learning-based model was verified to be capable of differentiating NMOSD from MS with an accuracy comparable to that of neurologists, exhibiting distinct advantages in terms of the consistency of classification. The proposed model has the potential to aid differential diagnosis of two important CNS inflammatory diseases in clinical practice. Further research is necessary to determine the applicability of this model in clinical settings and to determine whether the utilization of the model can lead to improved patient care and prognoses.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of the Asan Medical

Center. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

D-WK and K-KK contributed to conception and design of the study. HK and YL wrote the manuscript. Y-ML, E-JL, and K-KK collected clinical and MRI data. YO and HWK classified the MRI data. HK, YL, and JL conducted data analysis. D-WK and E-JL acquired funding. Y-ML, E-JL, Y-HK, JW, S-KJ, D-WK, and K-KK supervised the study and revised the manuscript. All authors

contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

1. Hoftberger R, Lassmann H. Inflammatory demyelinating diseases of the central nervous system. *Handbook Clin Neurol.* (2017) 145:263–83. doi: 10.1016/B978-0-12-802395-2.00019-5

2. 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) 364:2106–12. doi: 10.1016/S0140-6736(04)17551-X

3. Kim S-M, Kim S-J, Lee HJ, Kuroda H, Palace J, Fujihara K. Differential diagnosis of neuromyelitis optica spectrum disorders. *Ther Adv Neurol Disord.* (2017) 10:265–89. doi: 10.1177/1756285617709723

4. 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) 18:113–5. doi: 10.1177/1352458511431973

5. 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) 69:239–45. doi: 10.1001/archneurol.2011.216

6. 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) 67:1016–7. doi: 10.1001/archneurol.2010.188

7. 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) 84:1165–73. doi: 10.1212/WNL.0000000000001367

8. Waters PJ, Pittock SJ, Bennett JL, Jarius S, Weinshenker BG, Wingerchuk DM. Evaluation of aquaporin-4 antibody assays. *Clin Exp Neuroimmunol.* (2014) 5:290–303. doi: 10.1111/cen3.12107

9. Geraldes R, Ciccarelli O, Barkhof F, De Stefano N, Enzinger C, Filippi M, et al. The current role of MRI in differentiating multiple sclerosis from its imaging mimics. *Nat Rev Neurol.* (2018) 14:199–213. doi: 10.1038/nrneurol.2018.14

10. 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. *Am J Neuroradiol.* (2017) 38:1145–50. doi: 10.3174/ajnr.A5173

11. 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. doi: 10.1016/j.jneumeth.2013.11.016

12. Lee H, Lee EJ, Ham S, Lee HB, Lee JS, Kwon SU, et al. Machine learning approach to identify stroke within 4.5 hours. *Stroke.* (2020) 51:860–6. doi: 10.1161/STROKEAHA.119.027611

13. Eshaghi A, Riyahi-Alam S, Saeedi R, Roostaei T, Nazeri A, Aghsaei A, et al. Classification algorithms with multi-modal data fusion could accurately distinguish neuromyelitis optica from multiple sclerosis. *NeuroImage Clin.* (2015) 7:306–14. doi: 10.1016/j.nicl.2015.01.001

14. Eshaghi A, Wottschel V, Cortese R, Calabrese M, Sahraian MA, Thompson AJ, et al. Gray matter MRI differentiates neuromyelitis optica from multiple sclerosis using random forest. *Neurology.* (2016) 87:2463–70. doi: 10.1212/WNL.0000000000003395

15. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436. doi: 10.1038/nature14539

16. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the brain: artificial intelligence in stroke imaging. *J Stroke.* (2017) 19:277–85. doi: 10.5853/jos.2017.02054

17. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology.* (2017) 284:574–82. doi: 10.1148/radiol.2017162326

18. 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) 316:2402–10. doi: 10.1001/jama.2016.17216

19. Adams JG, Melhem ER. Clinical usefulness of T2-weighted fluid-attenuated inversion recovery MR imaging of the CNS. *Am J Roentgenol.* (1999) 172:529–36. doi: 10.2214/ajr.172.2.9930818

20. Gawne-Cain ML, O'Riordan JI, Thompson AJ, Moseley IF, Miller DH. Multiple sclerosis lesion detection in the brain: a comparison of fast fluid-attenuated inversion recovery and conventional T2-weighted dual spin echo. *Neurology.* (1997) 49:364–70. doi: 10.1212/WNL.49.2.364

21. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapp.* (2019) 40:4952–64. doi: 10.1002/hbm.24750

22. Schmidt P. *Bayesian Inference for Structured Additive Regression Models for Large-Scale Problems With Applications to Medical Imaging.* Munich: Ludwig-Maximilians-Universität München (2017).

23. Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM, editors. 3D statistical neuroanatomical models from 305 MRI volumes. In: *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference* San Francisco, CA (1993).

24. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *NeuroImage.* (2012) 62:782–90. doi: 10.1016/j.neuroimage.2011.09.015

25. Eitel F, Soehler E, Bellmann-Strobl J, Brandt AU, Ruprecht K, Giess RM, et al. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage Clin.* (2019) 24:102003. doi: 10.1016/j.nicl.2019.102003

26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Las Vegas, NV (2016). p. 770–7.

27. Cole JH, Poudel RPK, Tsagkrasoulis D, Caan MWA, Steves C, Spector TD, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage.* (2017) 163:115–24. doi: 10.1016/j.neuroimage.2017.07.059

28. Korolev S, Safiullin A, Belyaev M, Dodonova Y. Residual and plain convolutional neural networks for 3D brain MRI classification. In: *IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. Melbourne, VIC (2017). p. 835–8. doi: 10.1109/ISBI.2017.7950647

29. Zou L, Zheng J, Miao C, Mckeown M, Wang Z. 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. *IEEE Access* 5:23626–36. doi: 10.1109/ACCESS.2017.2762703

30. Yuan L, Wei X, Shen H, Zeng L, Hu D. Multi-center brain imaging classification using a novel 3D CNN approach. *IEEE Access* 6:49925–34. doi: 10.1109/ACCESS.2018.2868813

31. Xie S, Girshick RB, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI (2017). p. 5987–95. doi: 10.1109/CVPR.2017.634

32. Loshchilov I, Hutter F, editors. SGDR: stochastic gradient descent with warm restarts. In: *ICLR*. Toulon (2017).

33. Lin T-Y, Goyal P, Girshick RB, He K, Dollár P. Focal loss for dense object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice (2017). p. 2999–3007. doi: 10.1109/ICCV.2017.324

34. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. (1934) 26:404–13. doi: 10.1093/biomet/26.4.404

35. DeLong ER, Vernon WB, Bollinger RR. Sensitivity and specificity of a monitoring test. *Biometrics*. (1985) 41:947–58. doi: 10.2307/2530966

36. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology*. (2010) 73:1167–79. doi: 10.1016/j.theriogenology.2010.01.003

37. 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) 80:1330–7. doi: 10.1212/WNL.0b013e3182887957

38. 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) 88:132–6. doi: 10.1136/jnnp-2016-314005

39. Yoo Y, Tang LYW, Kim S-H, Kim HJ, Lee LE, Li DKB, et al., editors. *Hierarchical Multimodal Fusion of Deep-Learned Lesion and Tissue Integrity Features in Brain MRIs for Distinguishing Neuromyelitis Optica from Multiple Sclerosis*. Cham: Springer International Publishing (2017). doi: 10.1007/978-3-319-66179-7_55

40. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence vs. clinicians in disease diagnosis: systematic review. *JMIR Med Inform*. (2019) 7:e10010. doi: 10.2196/10010

41. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence vs. clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Bmj*. (2020) 368:m689. doi: 10.1136/bmj.m689

42. Kim SH, Mealy MA, Levy M, Schmidt F, Ruprecht K, Paul F, et al. Racial differences in neuromyelitis optica spectrum disorder. *Neurology*. (2018) 91:e2089-e99. doi: 10.1212/WNL.0000000000006574

43. Jarius S, Paul F, Franciotta D, Ruprecht K, Ringelstein M, Bergamaschi R, et al. Cerebrospinal fluid findings in aquaporin-4 antibody positive neuromyelitis optica: results from 211 lumbar punctures. *J Neurol Sci*. (2011) 306:82–90. doi: 10.1016/j.jns.2011.03.038