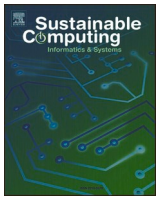
Sustainable Computing: Informatics and Systems 35 (2022) 100706

Contents lists available at ScienceDirect

Sustainable Computing: Informatics and Systems

journal homepage: www.elsevier.com/locate/suscom



A Convolutional Neural Network model for identifying Multiple Sclerosis on brain FLAIR MRI

Züleyha Yılmaz Acar [a,*], Fatih Başçiftçi [a], Ahmet Hakan Ekmekci [b]

[a] Selçuk University, Faculty of Technology, Department of Computer Engineering, Turkey
[b] Selçuk University, Faculty of Medicine, Department of Neurology, Turkey

ARTICLE INFO

Keywords:
Multiple Sclerosis
Convolutional Neural Network
Magnetic resonance imaging
Deep learning
Multiple sclerosis identification

ABSTRACT

Multiple Sclerosis (MS) is a neurodegenerative disease that occurs because of demyelination in nerve cells. Early treatment can be provided, and its progression can be prevented with an early diagnosis of the disease. The most remarkable finding in the identifying of MS disease is white matter lesions in the brain, which can be detected by magnetic resonance imaging (MRI). In this study, the identification of MS was performed with the proposed Convolutional Neural Network (CNN) model by detecting the presence of lesions from the brain Fluid-Attenuated Inversion Recovery (FLAIR) Magnetic Resonance (MR) images. The features of MS lesions in MR images are extracted with the proposed CNN model as an efficient and useful model with a low number of trainable parameters. The proposed CNN model has been compared with the traditional machine learning and state-of-the-art DL methods on a 5-fold cross-validation procedure. All methods are implemented on the same dataset. The results were obtained with both slice-level and patient-level data splitting methods. According to the results of slice-level splitting, the proposed CNN model achieved better success with the accuracy of 98.0% (± 0.02), the sensitivity of 97.9% (± 0.03), specificity of 98.3% (± 0.03), precision of 98.2% (± 0.03) values. In the results obtained with patient-level splitting, the accuracy of 90.3% (± 0.05), the sensitivity of 90.5% (± 0.05), the specificity of 90.1% (± 0.09), and the precision of 91.1% (± 0.09). The proposed CNN model obtained high and consistent performance in both splitting methods compared to other methods.

1. Introduction

Multiple Sclerosis (MS) is a neurodegenerative, chronic, and inflammatory central nervous system disease that occurs as a result of the demyelination of nerve cells, often seen in young adults [1]. Currently, Magnetic Resonance Imaging (MRI) is a widely-used imaging technique for the identification and observing of MS and determining the treatment process [2]. It is very important to distinguish MS in brain MRI findings from other demyelinating diseases (such as Schilder's Disease [3], Neuromyelitis Optica (NMO) [4], Acute Disseminated Encephalomyelitis (ADEM) [5]) and to correctly diagnose white matter lesions [6].

Additionally, the identification of MS lesions caused by the problems with MRI devices may not be fast and clear [7]. Early and accurate diagnosis of MS with MRI findings, which is recurrent and potentially crippling, is very important to minimize attacks with early treatment and to change the natural course of MS [8].

Identification of MS disease through MRI data has been a field of study for many years. Traditional Machine Learning (ML) methods were used first. These methods require manual preprocessing steps to the data. In these preprocesses, feature extraction and feature selection processes are applied. On the other hand, the Deep Learning (DL) models, whose popularity has increased in recent years, have become one of the frequently used methods for the identification of MS disease. They are very effective in representing the patterns of the disease with their multi-layered structure by eliminating the manual feature extraction steps.

In this study, a Convolutional Neural Network (CNN) model, one of the deep learning architectures, is proposed for the identification of MS disease from brain Fluid-Attenuated Inversion Recovery (FLAIR) MRI. MR images with FLAIR sequence are used as input images of the model. The model classifies the MR images as with or without lesions. The proposed model consists of a combination of convolution layers with Rectified Linear Unit (RELU), pooling layers, and batch normalization. Finally, it uses fully connected layers and a SoftMax activation function.

* Corresponding author.
E-mail address: zuleyhayilmaz@selcuk.edu.tr (Z. Yılmaz Acar).

https://doi.org/10.1016/j.suscom.2022.100706
Received 28 May 2021; Received in revised form 24 January 2022; Accepted 24 February 2022
Available online 28 February 2022
2210-5379/

Therewith, it is aimed to have a low calculation cost of the proposed model with a low number of trainable parameters. Thus, MS identification was made through MR images with a simple and at the same time effective model.

In order to evaluate the success of the proposed model, we compared the performances of traditional ML and DL methods on the same dataset. Performance was compared with Logistic Regression (LR), k-Nearest Neighbor (kNN), Decision Tree (DT), and Support Vector Machine (SVM) algorithms as traditional ML algorithms. DL models are considered in two ways: pre-trained and state-of-the-art models. VGG16, VGG19 [9], ResNet50, ResNet101 [10], and InceptionV3 [11] (which is a version of GoogleNet [12]) are used as pre-trained DL models, and state-of-the-art models have been selected in studies [6,13]. To our knowledge, there is no study in the literature on MS identification on the dataset in ourstudy. Therefore, all the methods were coded by the authors in order to make a fair comparison. The results were obtained under the same conditions with the same dataset, verified by 5-fold cross-validation method. Also, the results were reported according to both slice-level and patient-level data splitting methods. The success of the methods was presented using classification performance metrics that are accuracy, sensitivity, specificity, and precision.

2. Related works

In the literature, studies on MS disease can be grouped as follows: the diagnosing of MS disease [14–16], the segmentation of MS lesions [17–19], and predicting of conversion between MS types [20–22]. Until now, there are some studies in the literature that identify MS diseases through Magnetic Resonance (MR) images. It is seen that both traditional ML algorithms and DL models are used in these studies. MR images of MS patients are classified by studies using various ML algorithms, such as SVM [23,24], DT [23,25], kNN [23], LR [26], Random Forest (RF) [27] and Artificial Neural Network (ANN) [28]. These methods require manual feature extraction from data followed by feature selection steps. The selected features have varying effects on the results, especially for high-dimensional biomedical data.

On the other hand, DL models, which have been used increasingly in recent years, have very effective results on biomedical applications [29–32]. DL models replace traditional ML methods in many application areas with their ability to automatically extract high-level features [33]. Among DL architectures, CNN is the most utilized model in them with its multi-layered and stable structure [34]. It is the most widely used DL model to identify MS disease.

The study [6] diagnoses MS with a CNN model including parametric RELU. The model is a 10-layers architecture combined with dropout and parametric RELU. The study [35] proposes a CNN model using stochastic pooling. The model is a 14-layers architecture combined with batch normalization, dropout, and stochastic pooling. The study [15] diagnoses MS disease with a 3D CNN model based on the heatmap method. The model has been first trained with the dataset of Alzheimer patients, and the transfer learning principle of deep learning has been used. The study [36] makes it with the Graph Convolutional Network model, where they suggest classifying four sub-clinical profiles of MS. The study [13] proposes the CNN structure for the diagnosis of the disease. The fully connected layer is not used in the architecture that they have achieved with the combination of convolution, RELU, and pooling structure. In the study [37], CNN architecture, which is used as input for susceptibility-weighted imaging images, which contains radiological signal information in the diagnosis of the disease, has been proposed. The model consists of convolution, RELU, pooling and fully connected layers. In the study [38], the classification is done with spectral information of images with the Haar wavelet transform method. The proposed model is a combination of CNN architecture and two-dimensional discrete Haar wavelet transform.

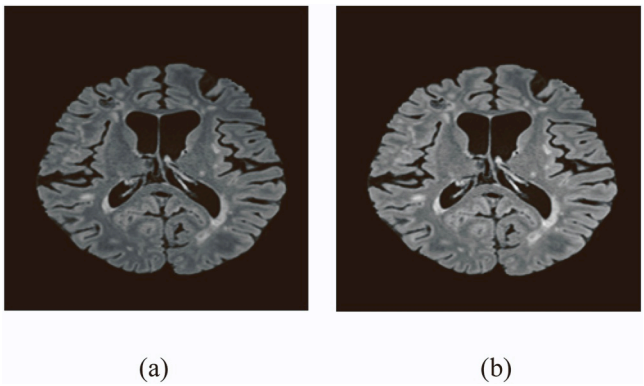


Fig. 1. The effect of HS normalization method on FLAIR MRI slice of patient01 in LabIT (a) original image (b) normalized image.

3. Materials and methods

3.1. Dataset

The dataset used in this study includes co-registered and bias-corrected brain MRI, brain mask, and ground truth data obtained by 3 T MR scanner for 30 MS patients with white matter lesion segmentation provided by the Laboratory of Imaging Technologies (LabIT) (publicly available at http://lit.fe.uni-lj.si/tools). Detailed information on the data can be found in [39]. There are two classes in this study: with MS lesion and without MS lesion. 971 FLAIR MRI slices with lesion and 593 FLAIR MRI slices without lesion were selected from the dataset. In order to overcome the data unbalanced problem, the oversampling method [40] was applied to without lesion class, which is the minority class. Since the images from the MR device do not always contain one-way images, it was sufficient to use the 180-degree rotation of randomly selected images from the minority class as a data augmentation method. As a result, a total of 1942 images, including 971 with lesions and 971 without lesions, were used in this study.

Since white matter lesions are one of the strongest findings of MS disease, we classified MRI slices as with or without lesions for the identification of MS disease in our study. We report the results both with patient-level data splitting of the dataset in train/test subsets (the most realistic scenario) and with slice-level data splitting, i.e., the same method used in [6,13,23,25–27,35,36,38], which is not realistic, but it is used for a direct comparison with the state-of-the-art models. We tested our model with the 5-fold cross-validation method to ensure the reliability of the results.

3.2. Preprocessing

In preprocessing step, firstly, skull-stripping and histogram stretching was applied to all MRI data. The skull-stripping was performed by overlapping the brain masks with the brain FLAIR MR images. Since the histogram distributions of the images are different from each other, the Histogram Stretching (HS) [41] process was performed on the skull-stripped images. HS method is an inter-scan intensity normalization method. The equation of HS is given with Eq. (1).

$$imOut(x,y) = (imIn(x,y) - imIn_{min})/(imIn_{max} - imIn_{min}) \quad (1)$$

Where $imOut(x,y)$ is the output normalized image pixel value, $imIn(x,y)$ is the original image pixel value; $imIn_{min}$ and $imIn_{max}$ is the minimum and maximum pixel value in the original image, respectively. The effect of the normalization process with the HS method with an example from with MS lesion class is illustrated in Fig. 1.

For traditional ML algorithms, feature extraction and feature selection preprocessing steps are applied to the skull-stripped and normalized images. As a feature selection step, a Histogram of Oriented Gradients

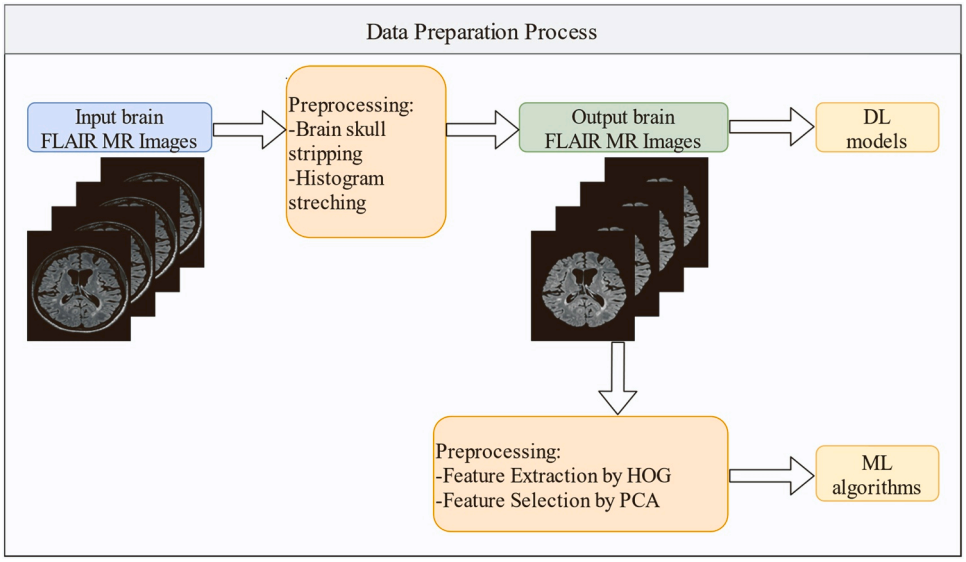


Fig. 2. The pipeline of the preprocessing steps.

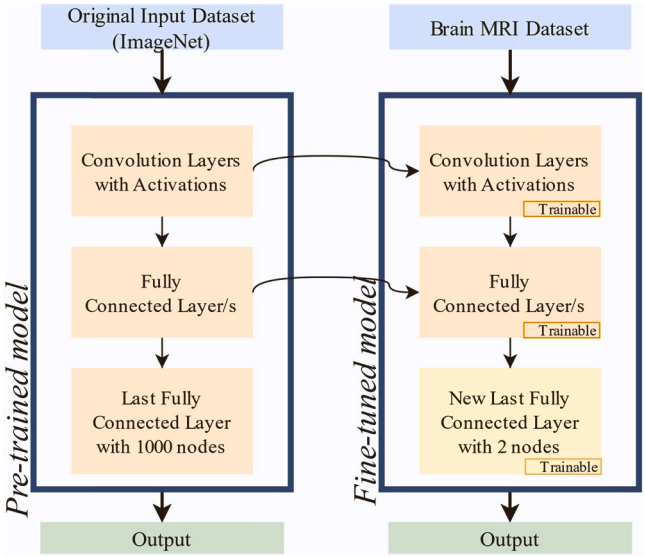


Fig. 3. Transfer learning strategy implemented in the study.

(HOG) [42] features of the MR images is utilized to generate the feature vectors in this study. HOG features are preferred due to performance and easy implementation [43]. Then, the Principal Component Analysis (PCA) method is applied to the feature vector. The PCA method has been widely applied as a feature selection method. The method is very effective and useful for reducing the number of features [44]. Therefore, the PCA method is implemented for dimensionality reduction and reduction of computation cost. The pipeline of the preprocessing steps is shown in Fig. 2.

3.3. Methodology

In this study, the classification process is carried out with both ML algorithms and DL models to identify MS diseases. ML algorithms are simple methods that classify feature vectors of the images [45]. Manual feature extraction and selection preprocessing steps are performed to generate input data in the ML algorithms.

We select the four well-known ML algorithms as a classifier in this study: SVM, LR, kNN, and DT. Besides, in recent years, successful results have been obtained with the application of deep learning techniques to biomedical data [46–49]. We proposed a model based on CNN for the identification of MS disease in this study. CNN is an important DL architecture consisting of a robust combination of multiple layers and is frequently used in computer vision applications [34]. We select the five

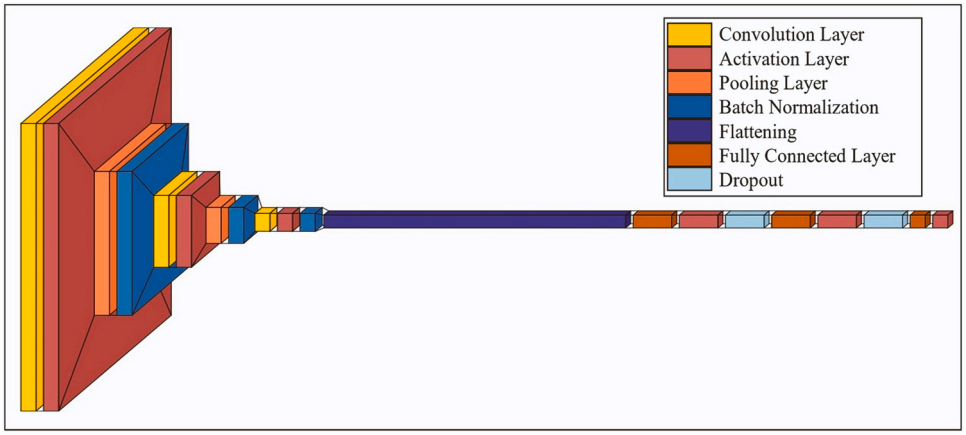


Fig. 4. Structure of the proposed CNN model.

Table 1
(a) Hyperparameters of the convolution layers of the proposed model. (b) Hyperparameters of the fully connected layers of the proposed model

(a)

Layers	Kernel size	# of filters	Stride	Padding
Convolution	(3,3)	32	2	Zero
MaxPooling	(2,2)	–	2	Zero
Convolution	(3,3)	64	2	Zero
MaxPooling	(2,2)	–	2	Zero
Convolution	(3,3)	128	2	Zero

(b)

Layers	# of nodes	Rate
Fully Connected	512	–
Dropout	–	0.5
Fully Connected	512	–
Dropout	–	0.5
Fully Connected	2	–

well-known pre-trained CNN models to compare the performance of the proposed model: VGG16, VGG19 [9], ResNet50, ResNet101 [10], and InceptionV3 [11]. In the meantime, the results have been compared with the performance of two state-of-the-art CNN models [6,13] proposed for MS identification in the literature.

The ML algorithms and state-of-the-art CNN models [6,13] are coded from scratch by the authors. Besides these, pre-trained models (i.e., VGG16, VGG19, ResNet50, ResNet101 and InceptionV3) are coded consistent with transfer learning strategy [50]. When applying this strategy, we changed the last fully connected layer with a new layer to modify the number of classes and re-trained the entire models with our dataset. The transfer learning strategy implemented to pre-trained DL models in this study is illustrated in Fig. 3.

4. The proposed CNN model

In this section, we described our CNN model to identify MS by using the presence of lesions from FLAIR brain MRI.

A general CNN model consists of convolution, activation, pooling and fully connected layers. The proposed model in this study includes three convolution layers, two pooling layers, and three fully connected layers. In addition, the dropout technique to prevent overfitting and batch normalization to achieve rapid convergence by normalizing the data, and also prevent overfitting has been integrated. The structure of the model is illustrated in Fig. 4.

This study focused on a simple and effective CNN model to identify MS disease. It is aimed to keep the number of parameters to be trained as low as possible. Thus, the computation cost and the risk of overfitting have been reduced. The hyperparameters were obtained by experimental studies. There is a comparison of the results obtained by different architectures in Sections 5.4 and 5.5.

The order of the layers described and hyperparameters of the model are listed in Table 1.

In our proposed CNN model, there are three convolution layers with a 3×3 filter size. Each convolution layer has 32, 64, and 128 filters with zero paddings respectively, and each layer is followed by the Rectified Linear Unit (RELU) activation layer to learning speed and classification accuracy [51].

After the convolution and activation layers, we used two maximum pooling layers with a 2×2 filter size to reduce the dimensionality and computational cost of the feature map.

Since a series of processes are performed on the layers in the DL models, the distribution of the data used can change across the DL layers. Batch normalization is utilized to normalize the data between these layers. The model is known to be effective in rapid convergence in the training phase [52]. This method is also used for regularization. In our model, batch normalization is utilized to normalize the distribution of the data and prevent overfitting problems.

Finally, there are three fully connected layers in the model. Each of the two fully connected layers has 512 nodes with dropout. The last fully connected layer has 2 nodes with the SoftMax activation function. All layers with input/output size of the proposed model are shown in Fig. 5.

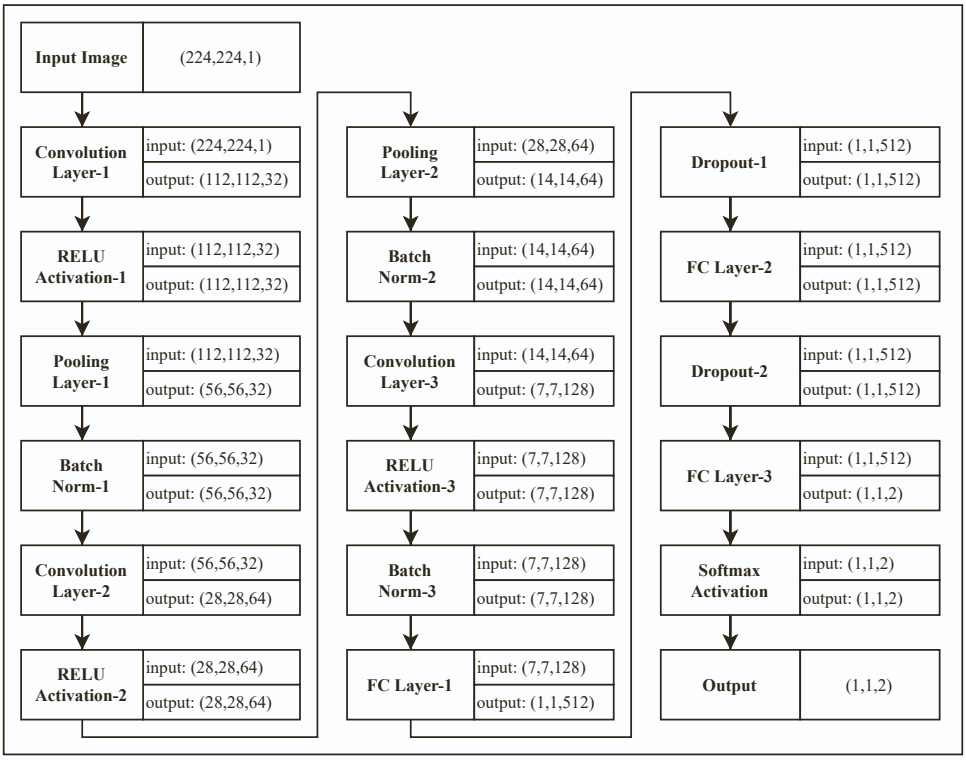


Fig. 5. The layers with input/output size of the proposed model (Norm: Normalization, FC: Fully Connected).

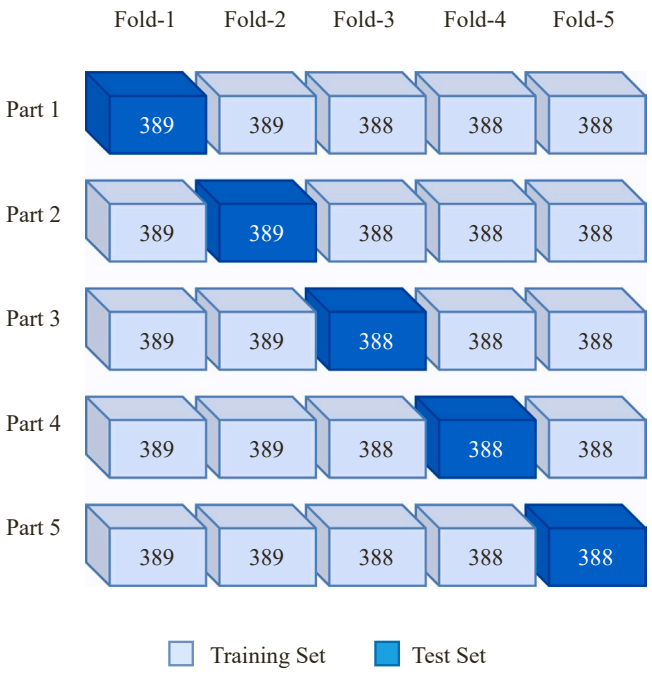


Fig. 6. The partitioning scheme of the data with 5-fold cross-validation.

5. Results and discussions

The performance of the ML and DL methods used for comparison in this study in the identification of MS disease was tested on brain FLAIR MR images. These all methods were implemented on Python using Keras [53] library. Details of ML algorithms are as follows: the kernel of the SVM algorithm is poly, the LR is regularized with liblinear solver, the number of features of the data is determined as 200 with HOG and PCA.

In DL models, VGG16, VGG19, ResNet50, ResNet101 and InceptionV3 are utilized as pre-trained models. The number of nodes in the last fully connected layers of the pre-trained models has been updated to be the number of classes in our problem. Re-training was carried out on these models with the dataset of this study. In all methods including the proposed model, the mini-batch size is set to 8 and the optimizer is Adam with learning late by 0.001. The number of epochs is set to 5 to avoid overfitting risk for slice-level splitting. In the patient-level splitting method, the number of epochs is set to 20 so that the training can be done sufficiently.

All results were obtained by K-fold cross-validation to evaluate the performance of the methods. According to the method, the data is divided into K-folds. K-1 folds are used for training data, one-fold is used for test data. The K-fold method was chosen because it is an effective method to show the advantage of the model [54]. The prediction success of the classifiers is the mean values obtained from K-folds. In this study, we fine-tuned the value of K = 5 in accordance with the dataset size and evaluated the performance of all methods with a 5-fold cross-validation process. The partitioning scheme of the data is shown in Fig. 6.

In this section, the results are reported under the same conditions, using both slice-level and patient-level data splitting methods. In the slice-level data splitting method, all slices of all patients are combined into a single dataset, then randomly separated into training and testing. We used this method to make direct comparisons with state-of-the-art models. On the other hand, in the patient-level data splitting method, it is guaranteed that all slices of a patient are included in only the training or only the test set. With this splitting method, more generalizable and realistic results are obtained since the models are tested with data that they have not encountered before.

Table 2
Comparison results of the proposed model with ML algorithms.

Method	ACC	SEN	SPE	PRE
SVM	0.925 (± 0.16)	0.932 (± 0.19)	0.914 (± 0.21)	0.898 (± 0.22)
LR	0.921 (± 0.16)	0.929 (± 0.20)	0.909 (± 0.21)	0.893 (± 0.22)
kNN	0.919 (± 0.17)	0.927 (± 0.20)	0.907 (± 0.22)	0.890 (± 0.23)
DT	0.914 (± 0.17)	0.922 (± 0.21)	0.901 (± 0.22)	0.883 (± 0.23)
Proposed CNN model	0.980 (± 0.02)	0.979 (± 0.03)	0.983 (± 0.03)	0.982 (± 0.03)

5.1. Performance metrics

To interpret the performance of all methods, we generate the confusion matrices of each method.

Thereby, the performances were compared with Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), and Precision (PRE) metrics. In this study, there are two classes: MR images with a lesion which is positive class, without lesion which is negative class. Accordingly, ACC indicates the correct prediction rate among the total number of predictions. SEN indicates the ability of the method to correctly identify the positive class. SPE indicates the ability of the method to correctly identify the negative class. PRE indicates the rate of correctly identification positive predictions to the total predicted positives. The equations of the metrics are expressed by Eqs. (2)–(5).

$$ACC = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

$$SEN = TP/(TP + FN) \quad (3)$$

$$SPE = TN/(TN + FP) \quad (4)$$

$$PRE = TP/(TP + FP) \quad (5)$$

Where *TP* (True Positive) refers to the number of positive predictions that are actually positive, *TN* (True Negative) refers to the number of negative predictions that are actually negative, *FP* (False Positive) refers to the number of positive predictions that are actually negative, *FN* (False Negative) value refers to the number of negative predictions that are actually positive.

5.2. According to slice-level data splitting

5.2.1. Comparison with ML algorithms

Firstly, we compare our proposed model and ML algorithms on classification performance metrics . The results are listed in Table 2.

According to Table 2, although the success is almost over 90%, it is seen that SVM achieves the highest success, and DT achieves the lowest

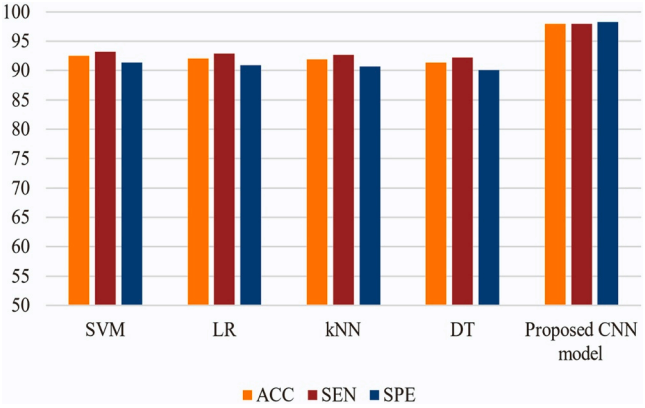


Fig. 7. Visual comparison of classification rates obtained by ML algorithms.

Table 3
Comparison results of the proposed model with pre-trained DL models.

Method	ACC	SEN	SPE	NTP (M)[a]
VGG16	0.778 (± 0.17)	0.912 (± 0.08)	0.655 (± 0.35)	~ 14.7 M
VGG19	0.912 (± 0.04)	0.910 (± 0.07)	0.911 (± 0.08)	~ 20.0 M
ResNet50	0.793 (± 0.18)	0.600 (± 0.35)	0.987 (± 0.02)	~ 23.6 M
ResNet101	0.663 (± 0.10)	0.534 (± 0.17)	0.786 (± 0.20)	~ 42.6 M
InceptionV3	0.844 (± 0.19)	0.812 (± 0.37)	0.878 (± 0.20)	~ 21.8 M
Proposed CNN model	0.980 (± 0.02)	0.979 (± 0.03)	0.983 (± 0.03)	~ 3.6 M

[a] M: millions.

Table 4
Comparison results of the proposed model with state-of-the-art CNN models.

Method	ACC	SEN	SPE	PRE
CNN Model-1	0.940 (± 0.05)	0.985 (± 0.02)	0.891 (± 0.11)	0.914 (± 0.08)
CNN Model-2	0.926 (± 0.04)	0.899 (± 0.09)	0.950 (± 0.08)	0.956 (± 0.06)
Proposed CNN model	0.980 (± 0.02)	0.979 (± 0.03)	0.983 (± 0.03)	0.982 (± 0.03)

success among ML algorithms. Besides all these, the highest results are obtained by a success of almost 98% with the proposed CNN model on ACC, SEN, SPE, and PRE metrics. The consistency of the results is expressed with the standard deviation values in Table 2. The proposed model reached the most robust results with the lowest standard deviation values compared to the ML algorithms.

The results show that ML algorithms cannot achieve sufficient success in recognizing the characteristics of MS lesions, which is the important finding defining MS disease. Manual feature extraction methods are insufficient. On the other hand, as a simple and effective the proposed model, defines the findings of MS lesions with a high level of performance. A visual comparison results of ACC, SEN, and SPE metrics as percentages are illustrated in Fig. 7.

5.2.2. Comparison with DL models

We evaluated the performances of the other DL models and proposed models on the same dataset under the same conditions. First, we compared the performance of our model with the pre-trained CNN models, which are VGG16, VGG19, ResNet50, ResNet101, and InceptionV3, on ACC, SEN, and SPE, and the number of trainable parameters (NTP) metrics. The results are listed in Table 3

As shown in Table 3, except for VGG19 and InceptionV3, other pre-trained DL models have an unbalanced distribution of SEN and SPE metrics. The ResNet50 and ResNet101 have quite a low SEN. They classify the MR images mostly as without lesions. This indicates that these models cannot characterize the features of MS lesions on the MR images, and this is a notable error in identifying MS. On the other hand, the VGG16 has quite a low SPE. It classifies the MR images mostly as with lesions. Classifying MR slices without lesions as with lesions may lead to incorrect estimation of the location of MS lesions in the brain and problems in understanding the characteristics of MS disease. Among the pre-trained DL models, significant results were obtained by VGG19 and InceptionV3. VGG19 outperforms other pre-trained DL models with ACC and balanced SEN and SPE distributions. However, it could not reach as high performance as our proposed CNN model. Although VGG19 achieves approximately 91% success, our model shows its superiority with an approximately 98% classification accuracy rate with balanced SEN and SPE distributions. When the standard deviation values of the results obtained with 5-fold cross-validation are examined, it is obvious that our

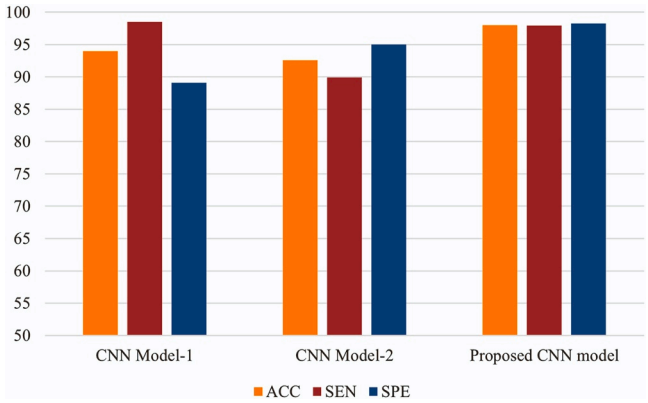


Fig. 8. Visual comparison of classification rates obtained by state-of-the-art CNN models.

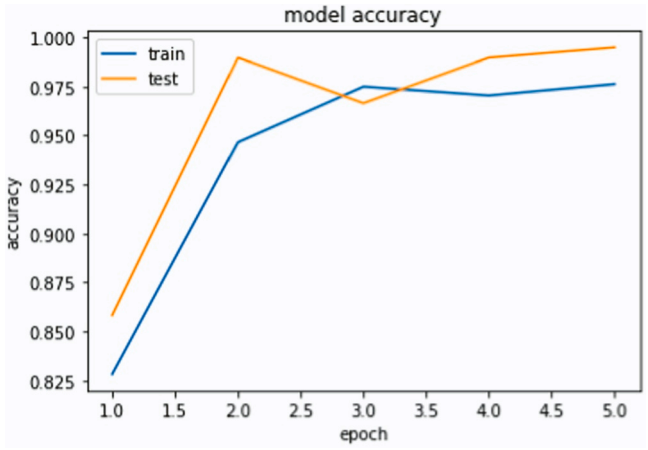


Fig. 9. The training and testing accuracy processes of the proposed CNN model.

model is the most stable and has generally obtained similar high-performance results for each fold.

Table 3 also contains the number of trainable parameters. The models other than ours have a quite number of parameters to be trained. This can cause more computation cost and overfitting, especially in datasets that are not too large. Our CNN model is the one with the lowest number of trainable parameters with ~3.6 M. The model recognizes very well the structure of MS lesions and achieves high classification rates owing to its efficiency.

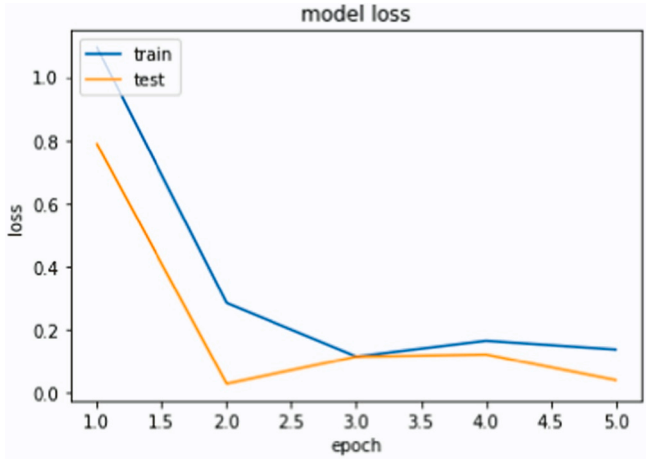


Fig. 10. The training and testing loss processes of the proposed CNN model.

Table 5
Comparison results of the proposed model with ML algorithms.

Method	ACC	SEN	SPE	PRE
SVM	0.589 (± 0.08)	0.588 (± 0.10)	0.385 (± 0.08)	0.562 (± 0.06)
LR	0.572 (± 0.08)	0.563 (± 0.15)	0.629 (± 0.12)	0.545 (± 0.18)
kNN	0.555 (± 0.07)	0.525 (± 0.12)	0.584 (± 0.09)	0.556 (± 0.08)
DT	0.530 (± 0.10)	0.415 (± 0.15)	0.650 (± 0.10)	0.537 (± 0.13)
Proposed CNN model	0.903 (± 0.05)	0.905 (± 0.05)	0.901 (± 0.09)	0.910 (± 0.09)

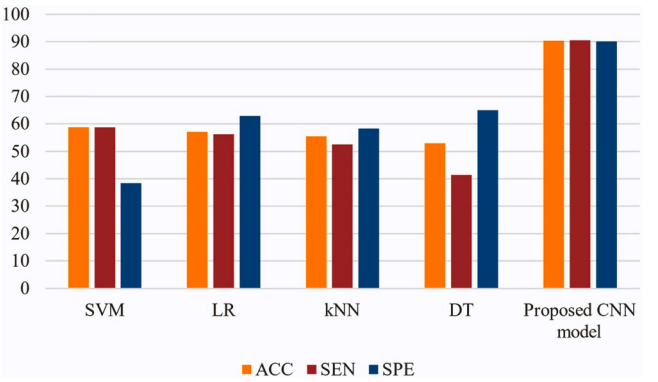


Fig. 11. Visual comparison of classification rates obtained by ML algorithms based on patient-level splitting.

Finally, the performance of our model was compared with the state-of-the-art CNN models and the results are given in Table 4.

The CNN models previously suggested for MS identification have been tested with the dataset of this study. The results show the superiority of our model over other state-of-the-art CNN models. Although the highest SEN value was obtained by CNN Model-1, the unbalanced SEN and SPE indicate that the features of MS lesions are not fully extracted. Besides all these, when analyzed the PRE metric, our model is significantly superior to the CNN Model-1 and Model-2. The PRE value is a measure that shows the rate of correct predictions in all classes and should be as high as possible. The PRE of our model is the highest value with 98% among other DL models. A visual comparison of models on ACC, SEN, and SPE metrics is illustrated in Fig. 8.

Additionally, the accuracy and loss process of the proposed CNN model during the training and testing process through the epochs are shown in Fig. 9 and Fig. 10, respectively.

When the accuracy and loss graphs in Fig. 9 and Fig. 10 are examined, it is seen that the model exceeded the 92.5% accuracy rate with its low loss value as of the second epoch with fast convergence.

Our model shows the success of distinguishing images with and without lesions with sufficient consistency by learning the lesion characteristics in brain MR images. In addition, the regular distribution of the train and test accuracy-loss graphs shows that the model is sufficiently trained with the training data without overfitting the problem and obtains the expected graphs over the test data.

5.3. According to patient-level data splitting

5.3.1. Comparison with ML algorithms

The results obtained with the patient-level data splitting method are listed in Table 5.

According to Table 5, the ML algorithms showed very low success when tested on data that they had not seen before. Although the SVM algorithm gave the best results, the accuracy remained at about 59%.

Table 6
Comparison results of the proposed model with pre-trained DL models.

Method	ACC	SEN	SPE
VGG16	0.507 (± 0.01)	0.800 (± 0.40)	0.200 (± 0.40)
VGG19	0.488 (± 0.01)	0.600 (± 0.49)	0.400 (± 0.49)
ResNet50	0.493 (± 0.01)	0.200 (± 0.40)	0.800 (± 0.40)
ResNet101	0.493 (± 0.01)	0.400 (± 0.49)	0.600 (± 0.49)
InceptionV3	0.533 (± 0.02)	0.067 (± 0.04)	1.000 (± 0.00)
Proposed CNN model	0.903 (± 0.05)	0.905 (± 0.05)	0.901 (± 0.09)

Table 7
Comparison results of the proposed model with state-of-the-art CNN models.

Method	ACC	SEN	SPE	PRE
CNN Model-1	0.870 (± 0.11)	0.825 (± 0.17)	0.913 (± 0.06)	0.893 (± 0.09)
CNN Model-2	0.854 (± 0.06)	0.832 (± 0.10)	0.873 (± 0.07)	0.872 (± 0.06)
Proposed CNN model	0.903 (± 0.05)	0.905 (± 0.05)	0.901 (± 0.09)	0.910 (± 0.09)

This ratio indicates that the learning is not adequately achieved, especially the characteristic features of MRI data without MS lesions. On the other hand, our CNN model has achieved sufficient success with balanced and consistent ACC, SEN and SPE values, exceeding 90%. A visual comparison results of ACC, SEN, and SPE metrics of the proposed model and ML algorithms as percentages are illustrated in Fig. 11.

5.3.2. Comparison with DL models

The results of pre-trained DL models obtained with patient-level data splitting are given in Table 6.

According to the patient-level splitting results in Table 6, none of the DL models could learn the characteristics of the MRI data. The models classified all data as either with or without MS lesions. We can confirm this information as the models have very high only SEN or only SPE values. With these results, it is concluded that pre-trained DL models cannot be successful in MRI data because they are pre-trained with ImageNet dataset.

On the other hand, in the results obtained with slice-level splitting, the VGG19 model showed 91% success, even though there was mostly an unbalance in the SEN and SPE values of the DL models. However, when the same model was run using the patient-level splitting method under the same conditions, it achieved the lowest ACC value of 49%. It

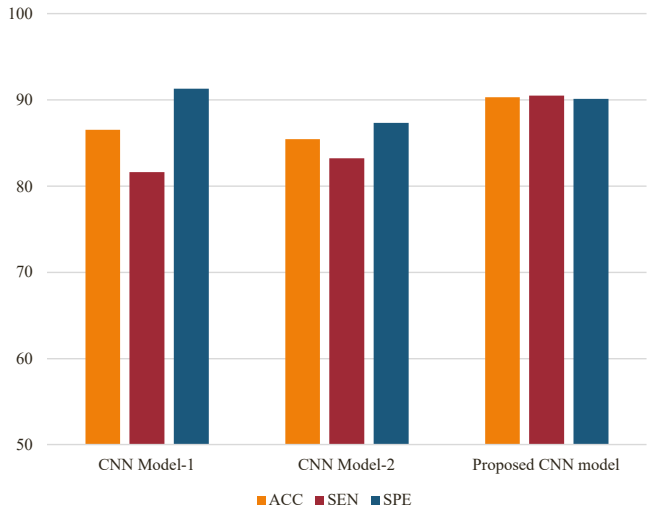


Fig. 12. Visual comparison of classification rates obtained by state-of-the-art CNN models based on patient-level splitting.

Table 8
Performance analysis (in %) of the CNN models with different numbers of convolution layers.

Number of convolution layers	Number of filters	ACC	SEN	SPE
2	128, 256	96.86	97.65	96.39
3	8,16,32	96.85	99.89	93.85
3 (Proposed)	32,64,128	**98.00**	97.94	98.28
4	32,64,128,256	96.35	95.72	96.89

Bold means the best.

suffered a great loss in success rate.

Another performance comparison was made between our proposed CNN model and state-of-the-art models with patient-level splitting. The results are listed in Table 7.

When Table 7 is examined, it is seen that the ACC values of CNN Model-1, Model-2 and our proposed model are 86.97%, 85.44%, and 90.33%, respectively. Although all results are higher than ML and DL methods, the best accuracy among them is achieved with the proposed CNN model. Together with these, while CNN Model-1 and Model-2 both have higher success in classifying MRI data without MS lesions, they have lower success in classifying MRI data with MS lesions. In other words, state-of-the-art CNN models were insufficient to learn MS lesion characteristics under the same conditions.

The proposed CNN model achieved a balanced success in learning the features of both classes, with a SEN value of 90.51% and an SPE value of 90.14%. Since the presence of MS lesions is one of the most important markers of MS disease, the result of the PRE value is very important. The proposed model achieved the highest result with a PRE value of 91.07% compared to state-of-the-art models. The low standard deviation values in all results mean that our proposed model has close classification success for each fold.

The visual comparison of the comparison is given in Fig. 12. The balanced and high results of the proposed model are shown in the figure.

Compared with the ML and pre-trained DL methods, the best results were obtained with the state-of-the-art and the proposed CNN models in the mean values obtained with 5-fold cross validation with patient-level data splitting. The ML and pre-trained DL methods have not been successful in learning the characteristic features of MRI data, and their classification success is very low on test data they have not encountered before. On the other hand, the problem-specific development and the creating of the model from scratch show that they can achieve a certain level of success, so that the models can be generalized. The proposed CNN model in the study shows that it is the most effective model on both ML, DL, and state-of-the-art methods. Our model has the highest performance success and produces the most stable results with the best standard deviation values.

5.4. Analysis of the proposed CNN model

The analysis of the proposed architecture is explained in this section. The proposed CNN model in this study was created by experimental studies. Different architectures were created using the different number of convolution layers and number of filters. The results were obtained with slice-level splitting.

The number of convolution layers was determined as 2, 3, and 4 layers. While creating the architectures, we used the number of filters starting from 8 and increasing by 2 times up to 256. All remaining hyperparameter values were kept the same while obtaining the results with these combinations.

Table 8 below shows the results of the architectures with the highest accuracy value from the models with different numbers of convolution layers.

According to Table 8, in the experiments using 2 convolution layers, the highest accuracy value was reached with the filter numbers 128 and 256. When the results are examined, it is seen that when 2 convolution layers are used, balanced SEN and SPE values but low ACC values are obtained. The model remained shallow to achieve high accuracy values.

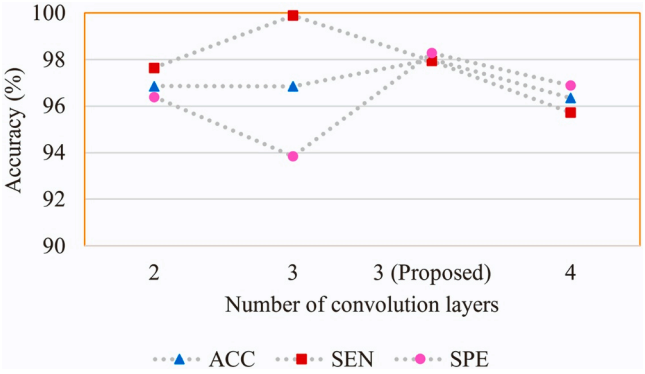


Fig. 13. The effect of different numbers of convolution layers on the performance.

Table 9
Comparing the results (in %) of CNN models with dropout and batch normalization.

Methods	ACC	SEN	SPE
Dropout	63.59	100	27.11
Without Dropout and Batch Normalization	95.93	96.65	95.22
1 Batch Normalization Layer	96.39	99.79	92.87
2 Batch Normalization Layers	93.71	87.13	99.36
3 Batch Normalization Layers (Proposed)	**98.00**	97.94	98.28

Bold means the best.

Except for the proposed model, the other architecture with 3 convolution layers that gives the highest ACC value is the model which has filter numbers of 8, 16, and 32. However, in this model, due to the low number of filters, sufficient training could not be achieved and unbalanced SEN and SPE values were obtained. According to this result, it is seen that most of the test data are classified as MR images with lesions and underfitting has occurred for this class. Therefore, a very high SEN value was obtained.

Finally, among the 4-convolution layer architecture, the combination with filter numbers 32, 64, 128, and 256 gave the highest ACC value. The results are relatively low results in spite of balanced SEN and SPE values. Increasing the number of filters or the deeper model did not improve the results.

The results are illustrated in Fig. 13.

As seen in Fig. 13, the most similar ACC, SEN and SPE values were obtained in the proposed model. The proposed model reached the highest ACC value with filter numbers of 32, 64, and 128. With the proposed model, both balanced SEN and SPE values and the highest ACC value among the different number of convolution layers were achieved. The proposed model is the best combination which analyzes the lesion characteristic in brain MR images with the highest ACC value of 98.0%, balanced SEN value of 97.94% and SPE value of 98.28%.

5.5. Effect of the batch normalization

Another experimental work was to apply the batch normalization process to our model.

The distribution of the data changes as mathematical operations are performed between the layers in deep learning [52,55]. This distribution needs to be regularized. In addition, studies with a small data set need to take precautions in case of overfitting. In our proposed model, we used the batch normalization process to regularize the distribution of the data and to prevent overfitting. Since dropout is generally used to prevent overfitting, we also included the results obtained when using dropout between layers in experimental studies. The effect of both methods on

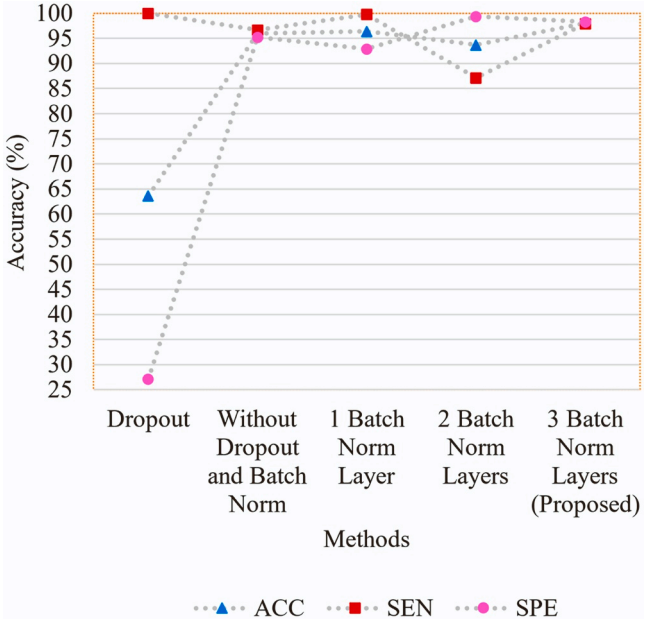


Fig. 14. The effect of dropout and batch normalization on performance in the proposed model (Norm: Normalization).

the results is given in Table 9.

According to Table 9, it is seen that the use of dropout between layers negatively affects the performance of the model to distinguish the features between classes. The model achieved an accuracy of 63.59%, a sensitivity of 100%, and a specificity of 27.11%. The high correlation between the convolution layers was interrupted due to the dropout method and the learning was not fully realized. The model classified all test data as MR images with lesions. Without dropout or batch normalization, although balanced results are obtained in the model, the ACC, SEN, and SPE values are relatively low.

The following steps in Fig. 5 applied in the models with 1, 2, and 3 batch normalization layers: only Batch Norm-1; Batch Norm-1 and Batch Norm-2; Batch Norm-1, Batch Norm-2, and Batch Norm-3, respectively.

When the results obtained are examined, it shows that the application of 1 or 2 batch normalization layers is not sufficient for the model to learn the features of both classes. Utilization of batch normalization layers after all convolution and activation process shows that high and balanced ACC, SEN, and SPE results are obtained with 98.00%, 97.94%, and 98.28% values. The graph of the performance metrics in Table 9 is illustrated in Fig. 14.

The expectation in Fig. 14 is that similar and high ACC, SEN, and SPE values in the results. In the model designed with dropout, the SEN and SPE results are quite far from each other. Among the models in which batch normalization was used, the proposed model outperforms to other models with both balanced and high-performance results. The results show that the proposed model with the batch normalization method used after each convolution layer regularizes the data distribution and prevents overfitting.

6. Conclusion

The well-known finding in the identification of MS diseases is white matter lesions. The presence, location, size, and spread of these MS lesions can be examined with magnetic resonance imaging. In our study, a CNN model was developed to identify MS on brain FLAIR MR images. We have presented a model that classifies an MR image as with or without lesion, ensuring simple and effective extraction of MS lesion features. Four traditional machine learning algorithms, five pre-trained CNN models, and two state-of-the-art CNN models have been tested on the same dataset and compared the performances with the proposed CNN model. The results were obtained with both slice-level and patient-level data splitting methods. The proposed model outperforms the other classification methods with high accuracy of 98.0% (± 0.02), balanced sensitivity of 97.9% (± 0.03), specificity of 98.3% (± 0.03), precision of 98.2% (± 0.03) values on slice-level splitting. In the results obtained with patient-level splitting, the accuracy of 90.3% (± 0.05), the sensitivity of 90.5% (± 0.05), the specificity of 90.1% (± 0.09), and the precision of 91.1% (± 0.09). The proposed CNN model obtained high and consistent performance in both splitting methods for learning MS lesion characteristics compared to other methods. The results show that the model can be utilized as an effective and usable model for MS identification.

CRediT authorship contribution statement

Züleyha Yılmaz Acar: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Fatih Başçiftçi:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Ahmet Hakan Ekmekci:** Formal analysis, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the Selçuk University Instructor Training Program Unit with Project no. 2018 – OYP – 031.

References

[1] P. Thakur, A. Mohammad, Y.R. Rastogi, R.V. Saini, A.K. Saini, Yoga as an intervention to manage multiple sclerosis symptoms, J. Ayurveda Integr. Med. (2019).
[2] E.C. Klawiter, Current and new directions in MRI in multiple sclerosis, Continuum 19 (2013) 1058–1073.
[3] S. Bacigaluppi, G. Polonara, M.L. Zavanone, R. Campanella, V. Branca, S.M. Gaini, et al., Schilder's disease: non-invasive diagnosis?: a case report and review, Neurol. Sci. 30 (2009) 421–430.
[4] S. Jarius, B. Wildemann, Aquaporin-4 antibodies (NMO-IgG) as a serological marker of neuromyelitis optica: a critical review of the literature, Brain Pathol. 23 (2013) 661–683.
[5] S. Tenembaum, T. Chitnis, J. Ness, J.S. Hahn, Acute disseminated encephalomyelitis, Neurology 68 (2007) S23–S36.
[6] Y.-D. Zhang, C. Pan, J. Sun, C. Tang, Multiple sclerosis identification by a convolutional neural network with dropout and parametric ReLU, J. Comput. Sci. 28 (2018) 1–10.
[7] L. Bonanno, N. Mammone, S. De Salvo, A. Bramanti, C. Rifici, E. Sessa, et al., Multiple sclerosis lesions detection by a hybrid watershed-clustering algorithm, Clin. Imaging 72 (2021) 162–167.
[8] L.A. Rolak, The diagnosis of multiple sclerosis, Neurol. Clin. 14 (1996) 27–43.
[9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–8.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision 2015, arXiv preprint arXiv:1512.00567, 2015.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[13] A. Soltani, S. Nasri, Improved algorithm for multiple sclerosis diagnosis in MRI using convolutional neural network, IET Image Process. 14 (2020) 4507–4512.
[14] A. Ion-Mărgineanu, G. Kocevar, C. Stamile, D.M. Sima, F. Durand-Dubief, S. Van Huffel, et al., Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features, Front. Neurosci. 11 (2017) 398.
[15] F. Eitel, E. Soehler, J. Bellmann-Strobl, A.U. Brandt, K. Ruprecht, R.M. Giess, et al., Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation, NeuroImage: Clin. 24 (2019), 102003.

[16] B. Insley, S. Rizvi, J. Cahill, J. Stone, C. Eickhoff, Deep neural net forecasting of multiple sclerosis disease severity (P3.2-013), Neurology 92 (2019) P3.2–013.
[17] R. McKinley, R. Wepfer, L. Grunder, F. Aschwanden, T. Fischer, C. Friedli, et al., Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence, NeuroImage: Clin. 25 (2020), 102104.
[18] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J.C. Vilanova, et al., Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach, NeuroImage 155 (2017) 159–168.
[19] N. Gessert, J. Krüger, R. Opfer, A.-C. Ostwaldt, P. Manogaran, H.H. Kitzler, et al., Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs, Comput. Med. Imaging Graph. 84 (2020), 101772.
[20] K. Bendfeldt, B. Taschler, L. Gaetano, P. Madoerin, P. Kuster, N. Mueller-Lenke, et al., MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry, Brain Imaging Behav. 13 (2019) 1361–1374.
[21] V. Mato-Abad, A. Labiano-Fontcuberta, S. Rodríguez-Yáñez, R. García-Vázquez, C. R. Munteanu, J. Andrade-Garda, et al., Classification of radiologically isolated syndrome and clinically isolated syndrome with machine-learning techniques, Eur. J. Neurol. 26 (2019) 1000–1005.
[22] H. Zhang, E. Alberts, V. Pongratz, M. Mühlau, C. Zimmer, B. Wiestler, et al., Predicting conversion from clinically isolated syndrome to multiple sclerosis–an imaging-based machine learning approach, Neuroimage Clin. 21 (2019), 101593.
[23] Y. Zhang, S. Lu, X. Zhou, M. Yang, L. Wu, B. Liu, et al., Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine, Simulation 92 (2016) 861–871.
[24] M. Weygandt, K. Hackmack, C. Pfüller, J. Bellmann–Strobl, F. Paul, F. Zipp, et al., MRI pattern recognition in multiple sclerosis normal-appearing brain areas, PLoS One 6 (2011), e21138.
[25] S. Jain, N. Rajpal, J. Yadav, Multiple Sclerosis Identification Based on Ensemble Machine Learning Technique, Available at SSRN 3734806, 2020.
[26] S. Wang, T. Zhan, Y. Chen, Y. Zhang, M. Yang, H. Lu, et al., Multiple sclerosis detection based on biorthogonal wavelet transform, RBF kernel principal component analysis, and logistic regression, IEEE Access 4 (2016) 7567–7576.
[27] D.R. Nayak, R. Dash, B. Majhi, Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests, Neurocomputing 177 (2016) 188–197.
[28] J. Han, S.-M. Hou, A multiple sclerosis recognition via hu moment invariant and artificial neural network trained by particle swarm optimization, in: Proceedings of the International Conference on Multimedia Technology and Enhanced Learning, 2020, pp. 254–64.
[29] P. Moeskops, M.A. Viergever, A.M. Mendrik, L.S. De Vries, M.J. Benders, I. Išgum, Automatic segmentation of MR brain images with a convolutional neural network, IEEE Trans. Med. Imaging 35 (2016) 1252–1261.
[30] S. Pereira, A. Pinto, V. Alves, C.A. Silva, Brain tumor segmentation using convolutional neural networks in MRI images, IEEE Trans. Med. Imaging 35 (2016) 1240–1251.
[31] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C.I. Sánchez, R. Mann, et al., Large scale deep learning for computer-aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.
[32] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, P.L. Carson, Medical breast ultrasound image segmentation by machine learning, Ultrasonics 91 (2019) 1–9.
[33] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: a new paradigm to machine learning, Arch. Comput. Methods Eng. (2019) 1–22.
[34] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: a review, Neurocomputing 187 (2016) 27–48.
[35] S.-H. Wang, C. Tang, J. Sun, J. Yang, C. Huang, P. Phillips, et al., Multiple sclerosis identification by a 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling, Front. Neurosci. 12 (2018) 818.
[36] A. Marzullo, G. Kocevar, C. Stamile, F. Durand-Dubief, G. Terracina, F. Calimeri, et al., Classification of multiple sclerosis clinical profiles via graph convolutional neural networks, Front. Neurosci. 13 (2019) 594.
[37] A. Lopatina, S. Ropele, R. Sibgatulin, J.R. Reichenbach, D. Güllmar, Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis, Front. Neurosci. 14 (2020).
[38] A. Alijamaat, A. NikravanShalmani, P. Bayat, Multiple sclerosis identification in brain MRI images using wavelet convolutional neural networks, Int. J. Imaging Syst. Technol. (2020).
[39] Ž. Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernuš, B. Likar, et al., A novel public MR image dataset of multiple sclerosis patients with lesion segmentation based on multi-rater consensus, Neuroinformatics 16 (2018) 51–63.
[40] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[41] C.P. Loizou, M. Pantziaris, I. Seimenis, C.S. Pattichis, Brain MR image normalization in texture analysis of multiple sclerosis, in: Proceedings of the 2009 9th International Conference on Information Technology and Applications in Biomedicine, 2009, pp. 1–5.
[42] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886–93.
[43] W. Zhou, S. Gao, L. Zhang, X. Lou, Histogram of oriented gradients feature extraction from raw Bayer pattern images, IEEE Trans. Circuits Syst. II: Express Briefs 67 (2020) 946–950.
[44] Q. Guo, W. Wu, D. Massart, C. Boucon, S. De Jong, Feature selection in principal component analysis of analytical data, Chemom. Intell. Lab. Syst. 61 (2002) 123–132.
[45] A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, Z. für Med. Phys. 29 (2019) 102–127.
[46] Y. Liu, P. Hao, P. Zhang, X. Xu, J. Wu, W. Chen, Dense convolutional binary-tree networks for lung nodule classification, IEEE Access 6 (2018) 49080–49088.
[47] H. Mohsen, E.-S.A. El-Dahshan, E.-S.M. El-Horbaty, A.-B.M. Salem, Classification using deep learning neural networks for brain tumors, Future Comput. Inf. J. 3 (2018) 68–71.
[48] C. Ge, Q. Qu, I.Y.-H. Gu, A.S. Jakola, Multi-stream multi-scale deep convolutional networks for Alzheimer's disease detection using MR images, Neurocomputing (2019).
[49] Y. Yoo, L.Y.W. Tang, D.K.B. Li, L. Metz, S. Kolind, A.L. Traboulsee, et al., Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from a clinically isolated syndrome, Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. (2017) 1–10.
[50] S. Khan, N. Islam, Z. Jan, I.U. Din, J.J.C. Rodrigues, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, Pattern Recognit. Lett. 125 (2019) 1–6.
[51] S. Pang, Z. Yu, M.A. Orgun, A novel end-to-end classifier using domain transferred deep convolutional neural networks for biomedical images, Comput. Methods Prog. Biomed. 140 (2017) 283–293.
[52] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, Int. Conf. Mach. Learn. (2015) 448–456.
[53] A. Gulli, S. Pal. Deep Learning with Keras, Packt Publishing Ltd, 2017.
[54] W. Zhang, C. Wu, Y. Li, L. Wang, P. Samui, Assessment of pile drivability using random forest regression and multivariate adaptive regression splines, Georisk: Assess. Manag. Risk Eng. Syst. Geohazards 15 (2021) 27–40.
[55] H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function, J. Stat. Plan. Inference 90 (2000) 227–244.