RESEARCH PAPER

DOI: 10.22080/frai.2026.30959.1045

Journal of Future Research in Artificial Intelligence and Internet of Things

Journal Home page: https://frai.journals.umz.ac.ir/

A Deep Learning–Driven Framework for Remote Monitoring and Severity Classification in Multiple Sclerosis Using Non-Motor Symptom Questionnaires

AliAsghar AkhavanMahdavi [1], Elham Mahdipour [2]*, MohammadAli Nahayati [3]

[1] *Computer Engineering Department, Khavaran Institute of Higher Education, Mashhad, Iran.*

[2] *Computer Engineering Department, Khavaran Institute of Higher Education.*

[3] *Faculty of Medicine, University of Medical Sciences, Mashhad, Iran, ORCID:0000-0001-8373-6053*

.

**Corresponding Author: elham.mahdipour@khavaran.ac.ir*

Article Info

Received 30 Dec 2025

Accepted 03 Feb 2026

Available online 12 Sep 2026

Keywords:

Multiple Sclerosis;

Deep Learning;

Patient-Reported Outcomes;

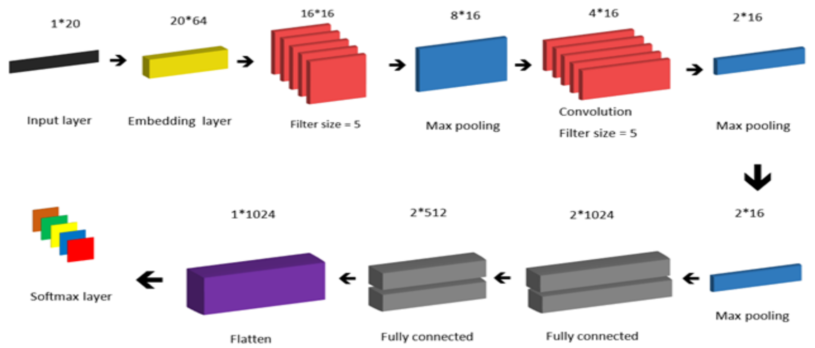
Remote Monitoring;

Telemedicine.

Abstract:

Multiple sclerosis (MS) is a chronic neurodegenerative disorder with unpredictable progression. Despite advancements in imaging and clinical assessment, access to regular neurological evaluations remains limited, particularly in regions with centralized healthcare services. This study proposes a novel deep learning–based approach for assessing MS severity using non-motor symptom questionnaires. Sixteen validated patient-reported outcome measures (PROMs) were administered through a web-based platform, enabling remote monitoring of emotional, behavioral, urinary, fatigue, and functional symptoms. Data from 152 fully completed questionnaires were used to train and evaluate the proposed convolutional neural network (CNN) model. The model demonstrated superior performance compared with traditional machine learning classifiers, achieving 99.5% accuracy on the test dataset while incorporating regularization to mitigate overfitting. The proposed framework demonstrates strong potential as a decision-support tool for assisting remote monitoring and early identification of symptom deterioration in patients with multiple sclerosis. The proposed approach offers an affordable, non-invasive, and scalable solution for MS monitoring, particularly beneficial where MRI and specialist access are limited. Future work will validate the model on larger, multicenter cohorts.

Graphical Abstract



© 2026 University of Mazandaran

Copyright © 2026 by the authors. Licensee FRAI, Babolsar, Mazandaran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (https://creativecommons.org/licenses/by/4.0/deed.en)

ISSN ***3092-7552***

Please cite this paper as:AA. AkhavanMahdavi, E. Mahdipour and MA. Nahayati, “A Deep Learning–Driven Framework for Remote Monitoring and Severity Classification in Multiple Sclerosis Using Non-Motor Symptom Questionnaires”, Future Research in AI and IOT, vol. 2, no. 2, pp. 65-77, 2026, doi: 10.22080/frai.2026.30959.1045.

1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease affecting the central nervous system and is a leading cause of neurological disability in young adults [1]. The global prevalence is estimated at approximately two million individuals, with epidemiological studies showing increasing incidence, particularly in women [2–4]. In Iran, prevalence varies significantly across regions, with documented rates reaching up to 74 per 100,000 people [3].

Clinical progression of MS is highly heterogeneous. Identifying disease severity and progression early is crucial to determining timely treatment strategies, preventing relapse, and improving quality of life [5–7]. Despite MRI being the gold standard for diagnosing and tracking neurological deterioration, regular imaging is costly and often inaccessible. Current wait times for neurologist appointments further exacerbate delayed care, with reports showing median delays exceeding 50 days [8–9].

Remote monitoring based on standardized patient-reported outcomes (PROMs) is a promising approach for early assessment of functional and psychological symptoms. Prior studies have highlighted the value of questionnaires in tracking fatigue, depression, anxiety, urinary dysfunction, and quality of life [10]. However, existing research typically focuses on (i) single questionnaire use, (ii) symptom-specific assessments, or (iii) binary classification.

1.1. Research Gap

There is no prior framework integrating multiple validated non-motor questionnaires with deep learning to classify multi-stage MS severity.

1.2. Objective

This research introduces a deep learning model capable of mapping PROM-derived feature vectors to five clinically established MS severity levels securely and remotely.

This study designed all questionnaires on a web-based platform, enabling patients to respond to questions at any time using any smartphone, tablet, or computer. Section 2 provides an overview of the relevant literature. Section 3 briefly discusses the standardized questionnaires used in this study. Section 4 presents the proposed deep learning-based method. Section 5 examines the experimental results of the proposed method and compares it with other machine learning algorithms. Section 6 discusses the advantages and disadvantages of the proposed method for monitoring multiple sclerosis from the perspective of specialized physicians in this field. Finally, Section7 presents the conclusions drawn from this study.

2. Related Works

Recent advances in artificial intelligence and digital health have provided new opportunities for monitoring disease progression in multiple sclerosis (MS). Existing studies in this domain can be categorized into four primary research streams: (1) deep learning and machine learning models for MS severity prediction, (2) remote and smartphone-based monitoring, (3) questionnaire-based assessments of non-motor symptoms, and (4) systematic reviews synthesizing AI applications in MS.

2.1. Deep Learning-based Severity Prediction

Several studies have proposed data-driven models to predict MS progression using clinical and imaging data. Zhang et al. [11] developed a multimodal deep neural network that combines electronic health records, clinical notes, and neuroimaging to estimate disease severity, demonstrating significant improvements in AUROC compared with unimodal approaches. Similarly, the MS Mosaic project [12] evaluated the performance of machine learning models in predicting high-severity symptoms and highlighted the utility of longitudinal self-reported data for remote monitoring. These findings suggest that deep learning can effectively capture complex patterns associated with disease progression; however, most existing approaches rely on specialized clinical data rather than accessible patient-reported outcomes.

2.2. Remote and Smartphone-Derived Digital Biomarkers

With increasing adoption of telehealth, several efforts have explored smartphone-based monitoring. Schwab and Karlen [13] introduced a deep learning model using smartphone sensor and interaction data to differentiate between MS and non-MS subjects, achieving promising classification performance. More recently, keystroke dynamics analysis was applied to passively estimate symptom severity, illustrating the feasibility of everyday digital interactions as remote biomarkers [14]. Nevertheless, these studies typically focus on motor or cognitive indicators and do not incorporate standardized non-motor symptom questionnaires.

2.3. Questionnaire-based and self-Report Assessments

Self-reported measures have long been recognized for their ability to capture psychological, cognitive, and fatigue-related symptoms that may not be immediately observable through imaging or clinical examination. For example, Rasmussen et al. [15] analyzed data from a large Danish self-report survey. They demonstrated the strong association between hidden symptoms such as fatigue, depression, and sleep disturbance with disability and quality of life. While these studies validate the importance of non-motor symptoms, they rarely employ machine learning or deep learning to translate questionnaire responses into automated multi-class severity predictions.

2.4. Systematic Reviews and Research Gap:

A recent systematic review by Vázquez-Marrufo et al. [16] categorized AI applications in MS into disease detection, relapse prediction, subtype classification, and progression analysis. The review highlighted the lack of research leveraging validated patient-reported outcome measures (PROMs) for automated severity classification using deep learning.

2.5. Summary and Gap

Although prior studies have demonstrated the potential of ML/DL models, smartphone-based monitoring, and questionnaire-based symptom evaluation, no existing work integrates multiple validated non-motor PROMs with a deep learning model to classify MS severity across five clinically accepted stages. Therefore, there remains a clear research gap in developing scalable and low-cost monitoring frameworks that leverage standardized patient-reported outcomes for automated severity assessment.

3. Questionnaires and Statistical Population (Case Study)

In this research, standard questionnaires were obtained from fellowship-trained physicians specializing in MS at Qaem Hospital in Mashhad. The study utilized 16 questionnaires, including the workplace questionnaire [17], anxiety and depression questionnaire [18], urinary issues questionnaire [19], general health questionnaire [20], quality of life questionnaire [21], fatigue severity questionnaire [22], sexual problems questionnaires [23], [24], the MFIS questionnaire [25], walking difficulties questionnaire [26], sleep quality questionnaire [27], hand function questionnaire [7], PARADISE questionnaire [28], FES questionnaire [29], medication adherence questionnaire [30], social integration questionnaire [31], and disease perception questionnaire [32].

All of these questionnaires were sourced from reputable articles, and their reliability and validity had been previously established.

Figure 1 shows the Persian-language web page designed for these questionnaires. Users could interact with the questionnaires by clicking on them and entering their responses. After users submitted their responses for each questionnaire, scores were automatically calculated according to predefined standards and stored in the corresponding database.

This observational study involved anonymous and voluntary participation. By completing the questionnaires, participants implicitly consented to the use of their anonymized data for research purposes. No identifiable information was collected, and the study adhered to general ethical principles and data privacy standards.

The study population consisted of 470 MS patients from Qaem Hospital in Mashhad. However, of the initial 470 patients, only 152 provided complete responses to the questionnaires. As a result, the statistical population of this research comprised 152 MS patients who had fully completed all or more than 10 of the relevant questionnaires.

The anxiety and depression questionnaire was found to have two sections, and the MFIS questionnaire had three subscales. Therefore, in this study, scores for each section were evaluated separately. This resulted in a CSV file with 21 features, which would be used as input to machine learning and deep learning algorithms. Each questionnaire score was treated as a feature. This resulted in a CSV file with 20 features or columns representing questionnaire scores and one column indicating the target class label, reflecting each patient's MS level. The levels of MS were defined based on discussions with experts and a review of relevant literature, categorizing patients into five levels of progression [33]: Level one represented clinically isolated syndrome (CIS), level two denoted relapsing-remitting (RR), level three indicated primary progressive (PP), level four represented secondary progressive (SP), and level five signified relapse-progressive (RP). Disease levels were encoded from 0 to 4 during implementation. Although EDSS is a widely used clinical scale, it relies heavily on neurological examination and motor function assessment. Since the proposed framework focuses on non-motor symptoms derived from patient-reported questionnaires, severity levels were defined through expert consensus to reflect multidimensional symptom burden. This approach enables scalable remote monitoring while acknowledging that EDSS-based validation remains an important direction for future work. So, patients were categorized into five clinical stages of MS, labeled from 0 to 4, corresponding to the following subtypes:

- Class 0: Clinically Isolated Syndrome (CIS) – 13 patients
- Class 1: Relapsing-Remitting (RR) – 66 patients
- Class 2: Primary Progressive (PP) – 45 patients
- Class 3: Secondary Progressive (SP) – 15 patients
- Class 4: Relapsing Progressive (RP) – 17 patients

Figure 2 displays a correlation matrix of the questionnaire scores and MS patient levels. We found no significant correlation between the features and the level of MS. Therefore, we used all the features to predict MS levels.

HTML, CSS, PHP, and a MySQL database were used to design the web pages, and Python was used to implement the machine learning algorithms and the proposed deep learning model.

4. Proposed Method

Machine learning is closely related to statistics and computer science, with a focus on using computers to make predictions and identify patterns. It is sometimes integrated with data mining, which focuses on exploratory data analysis. Machine learning can be used to learn and understand the underlying behavior of different entities and identify significant anomalies [34].

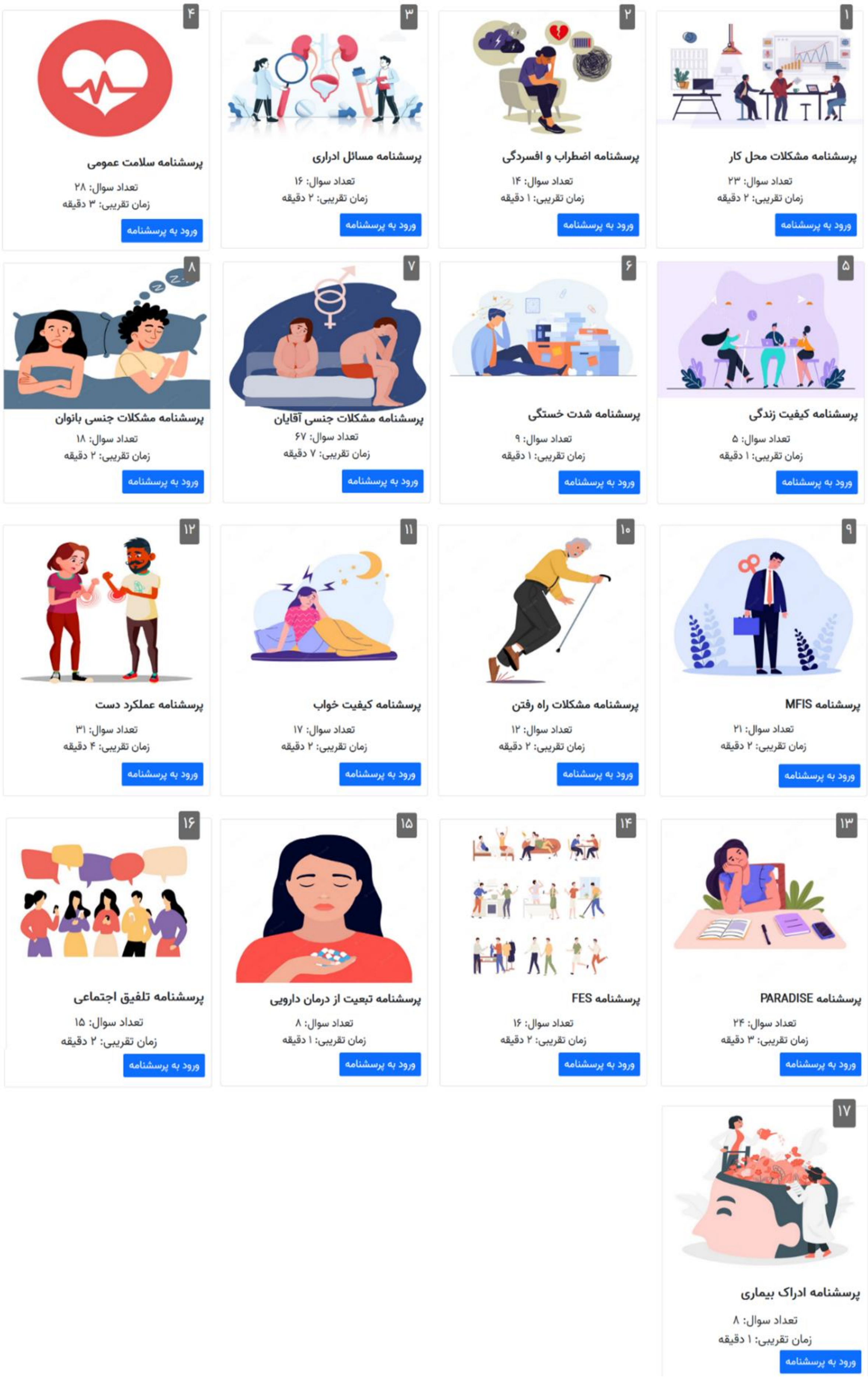


Figure 1. Web page designed for MS questionnaires in Persian

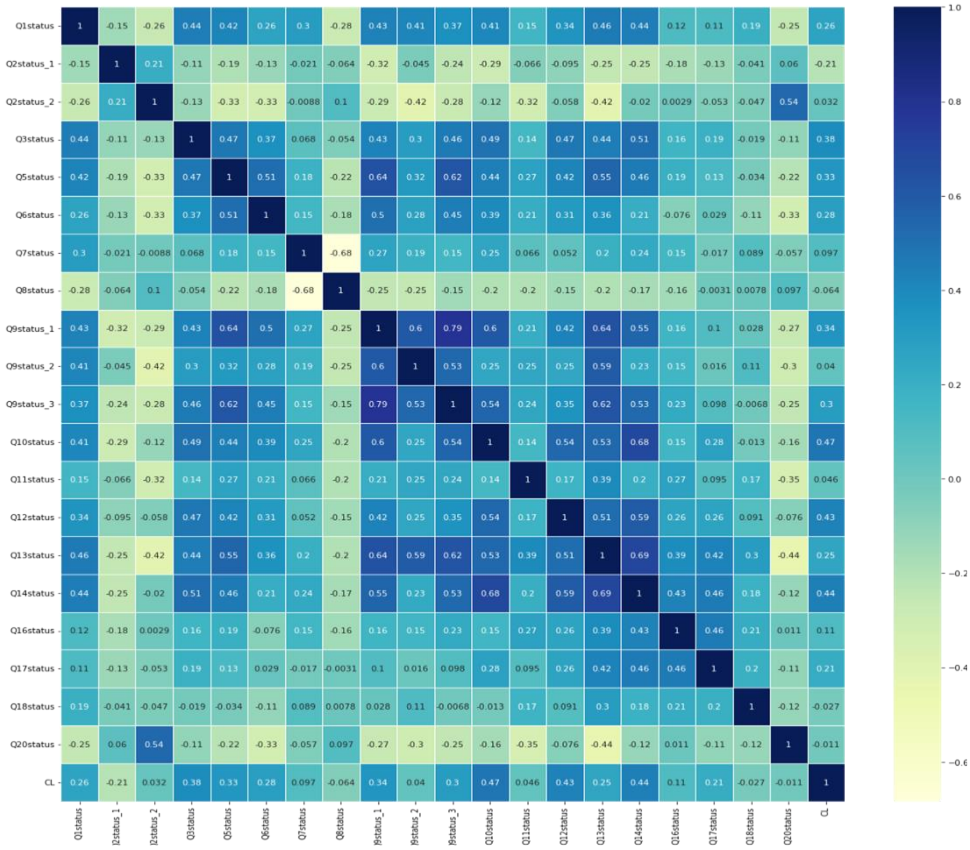


Figure 2. Heatmap of questionnaire scores and MS patient levels

In this study, we used machine learning to discover the relationship between patients' overall questionnaire responses and their MS levels. We proposed a deep learning-based model to identify patients' MS level, and we compared its performance to other machine learning algorithms, including decision trees (DT), naive Bayes (NB), k-nearest neighbors (KNN), and support vector machines (SVM). The innovations of this research include:

- Integrating 16 validated non-motor PROMs.
- Mapping them to five MS severity levels.
- Using a CNN model.
- Enabling low-cost remote monitoring.

Deep learning is a subset of machine learning that uses algorithms to learn high-level abstract concepts from data. This is done using a deep neural network, which is a type of artificial neural network with multiple layers. In other words, deep learning models learn knowledge and features from the data by processing it through multiple layers.

4.1. Questionnaire Scoring

As mentioned in Section 3, we used 16 standardized questionnaires, each with publicly available scoring methods described in their original references. The reliability and validity of all the employed questionnaires have been previously established. Based on patients' responses, we calculated the score for each questionnaire. Notably, the Hospital Anxiety and Depression Scale (HADS) [18] provides two separate scores—one for anxiety and one for depression—and the Modified Fatigue Impact Scale (MFIS) [25] yields three subscale scores relevant to clinical decision-making. Therefore, in the preprocessing phase for deep learning, we utilized a total of 20 questionnaire-derived scores, which were considered as input neurons in the neural network architecture.

4.2. Proposed Deep Learning Model

The proposed model is a basic convolutional neural network (CNN) architecture designed for multi-class classification. It takes integer-encoded inputs of length 20 and outputs probabilities for five different classes.

The model consists of six layers, as described below. The scores from 20 questionnaires are first input into the deep neural network. An embedding layer is used to increase the dimensionality of the input data. The architecture then includes two convolutional layers, two fully connected layers, a flatten layer, and an output layer.

Figure 3 illustrates the architecture of the proposed deep learning model, including the number of filters in each convolutional layer. The model first uses an embedding layer to encode the input—a sequence of 20 integers representing the questionnaire scores—into dense vectors of fixed size. In this case, the embedding layer takes an input of length 20 and produces an output of size 20×64.

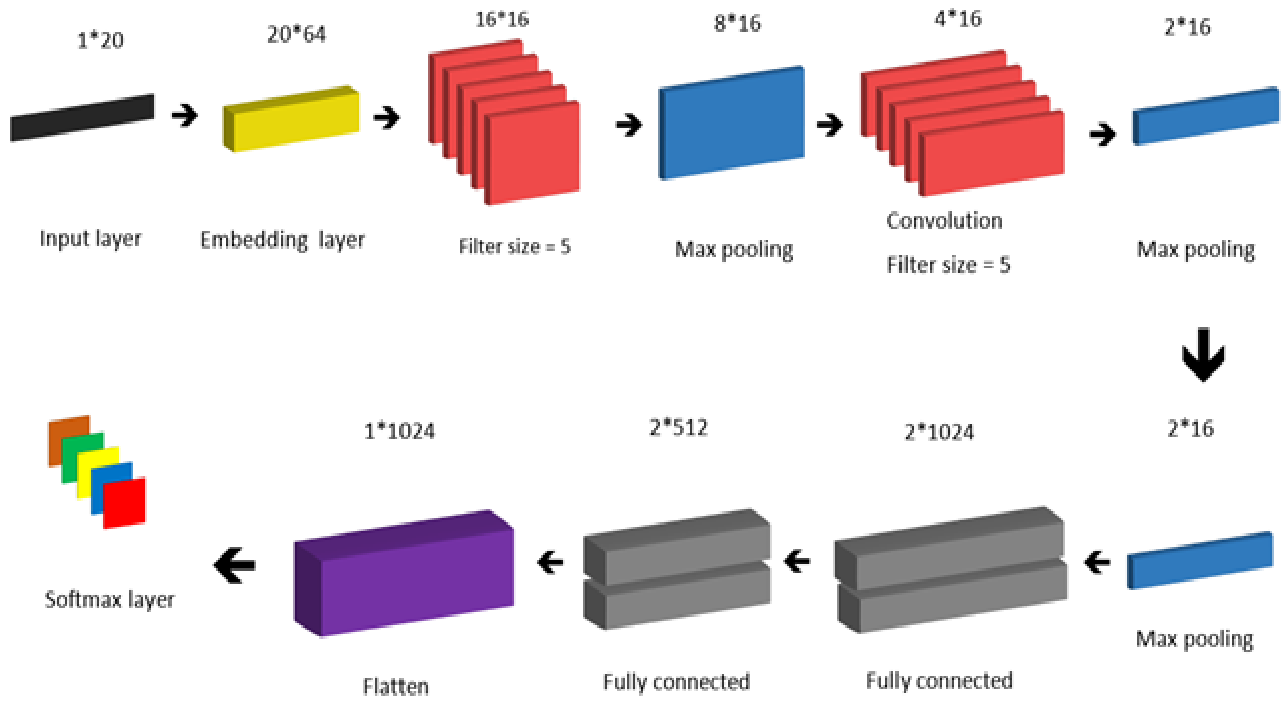


Figure 3. The proposed deep learning model

Next, a 1D convolutional layer with 16 filters of size 5 is applied, using the ReLU activation function. A 1D convolution is used because the input data is sequential. After the convolution, a max pooling layer is applied, reducing the spatial dimensions to 8×16.

Another 1D convolutional layer with a similar configuration is then added, followed by another max pooling layer. Afterward, two fully connected (dense) layers are included, with 1024 and 512 neurons, respectively.

Next, a flatten layer transforms the output tensor into a one-dimensional vector. Finally, the output layer uses a softmax activation function to predict the probabilities of the five MS levels.

Also, we used the following hyperparameters for tuning the proposed model:

- Optimizer: Adam
- Learning rate: 0.001
- Loss: categorical cross-entropy
- Batch size: 32
- Regularization: dropout 0.3, early stopping.

5. Experimental Results

To evaluate the performance of the proposed model, we used several standard classification metrics, including accuracy, precision, recall (sensitivity), F1-score, and the confusion matrix. Accuracy represents the proportion of correctly classified instances over the total number of instances. Precision is the ratio of true positives to the total predicted positives, indicating the model's ability to avoid false positives. Recall (or sensitivity) is the ratio of true positives to all actual positives, measuring how well the model captures relevant cases. F1-score is the harmonic mean of precision and recall, providing a balanced measure of both. The confusion matrix presents the distribution of true positives, false positives, false negatives, and true negatives across different classes, offering a detailed view of classification performance. All metrics were computed for each class and averaged where appropriate, using the scikit-learn library in Python.

As shown in Table 1, the proposed model outperforms conventional classifiers such as SVM and KNN across all key evaluation metrics. The dataset was split into 67% for training and 33% for testing across all methods. For the deep learning approach, 10% of the training data was further used as a validation set during training. The deep learning model required 1000 epochs to achieve the desired accuracy.

Note that the results in Table 1 are based on at least 20 distinct runs that accounted for data distribution. Similar results were obtained across all runs.

Figures 4 and 5 display the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) plots for the training and test sets.

Table 1. **Comparative performance of the proposed deep learning model and traditional machine learning classifiers using multiple evaluation metrics**

Train/Test	Algorithm	Precision	Recall	F1-score	Accuracy	Confusion Matrix
Train	**Deep Learning Proposed Method**	0.998	0.997	0.998	0.999	$\begin{bmatrix} 6 & 0 & 0 & 0 & 0 \\ 0 & 36 & 0 & 0 & 0 \\ 0 & 0 & 38 & 0 & 0 \\ 0 & 0 & 0 & 11 & 0 \\ 0 & 0 & 0 & 0 & 10 \end{bmatrix}$
	SVM	0.948	0.945	0.946	0.951	$\begin{bmatrix} 6 & 0 & 0 & 0 & 0 \\ 1 & 34 & 1 & 0 & 0 \\ 0 & 1 & 36 & 1 & 0 \\ 0 & 0 & 1 & 10 & 0 \\ 0 & 0 & 0 & 1 & 9 \end{bmatrix}$
	KNN	0.926	0.926	0.928	0.934	$\begin{bmatrix} 6 & 0 & 0 & 0 & 0 \\ 1 & 33 & 2 & 0 & 0 \\ 0 & 2 & 35 & 1 & 0 \\ 0 & 1 & 1 & 9 & 0 \\ 0 & 0 & 1 & 0 & 9 \end{bmatrix}$
	NB	0.872	0.865	0.867	0.88	$\begin{bmatrix} 5 & 1 & 0 & 0 & 0 \\ 3 & 30 & 3 & 0 & 0 \\ 2 & 2 & 31 & 3 & 0 \\ 0 & 1 & 2 & 8 & 0 \\ 0 & 1 & 1 & 1 & 7 \end{bmatrix}$
	DT	0.905	0.901	0.902	0.913	$\begin{bmatrix} 5 & 1 & 0 & 0 & 0 \\ 2 & 32 & 2 & 0 & 0 \\ 1 & 1 & 34 & 2 & 0 \\ 0 & 1 & 1 & 9 & 0 \\ 0 & 1 & 1 & 1 & 7 \end{bmatrix}$
Test	**Deep Learning Proposed Method**	0.995	0.994	0.995	0.997	$\begin{bmatrix} 7 & 0 & 0 & 0 & 0 \\ 0 & 29 & 1 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 7 \end{bmatrix}$
	SVM	0.918	0.915	0.916	0.924	$\begin{bmatrix} 6 & 1 & 0 & 0 & 0 \\ 0 & 27 & 2 & 1 & 0 \\ 0 & 1 & 6 & 0 & 0 \\ 0 & 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 & 6 \end{bmatrix}$
	KNN	0.893	0.89	0.891	0.902	$\begin{bmatrix} 6 & 1 & 0 & 0 & 0 \\ 1 & 26 & 2 & 1 & 0 \\ 0 & 1 & 6 & 0 & 0 \\ 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 1 & 0 & 6 \end{bmatrix}$
	NB	0.824	0.819	0.821	0.836	$\begin{bmatrix} 5 & 2 & 0 & 0 & 0 \\ 2 & 23 & 3 & 2 & 0 \\ 1 & 1 & 4 & 1 & 0 \\ 0 & 1 & 1 & 2 & 0 \\ 0 & 1 & 1 & 1 & 4 \end{bmatrix}$
	DT	0.871	0.865	0.868	0.883	$\begin{bmatrix} 6 & 1 & 0 & 0 & 0 \\ 1 & 25 & 3 & 1 & 0 \\ 0 & 1 & 5 & 1 & 0 \\ 0 & 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 1 & 5 \end{bmatrix}$

5.1. Robustness and Cross-Validation Analysis

To further evaluate the robustness and generalization capability of the proposed CNN model, an additional 5-fold cross-validation analysis was conducted. This analysis was performed as a complementary evaluation to the main train–test experimental setup. The dataset was randomly divided into five folds of approximately equal size. In each iteration, four folds were used for training and one fold for testing. The proposed model achieved a mean accuracy of 99.68% ± 0.08 and a mean F1-score of 99.48% ± 0.08 across the five folds. The small standard deviation (Std) indicates stable performance and limited sensitivity to data partitioning. Table 2 shows the five-fold cross-validation results.

Figure 6 illustrates the boxplot distribution of Accuracy and F1-score across the five folds, while Figure 7 presents the mean and standard deviation of the evaluated metrics.

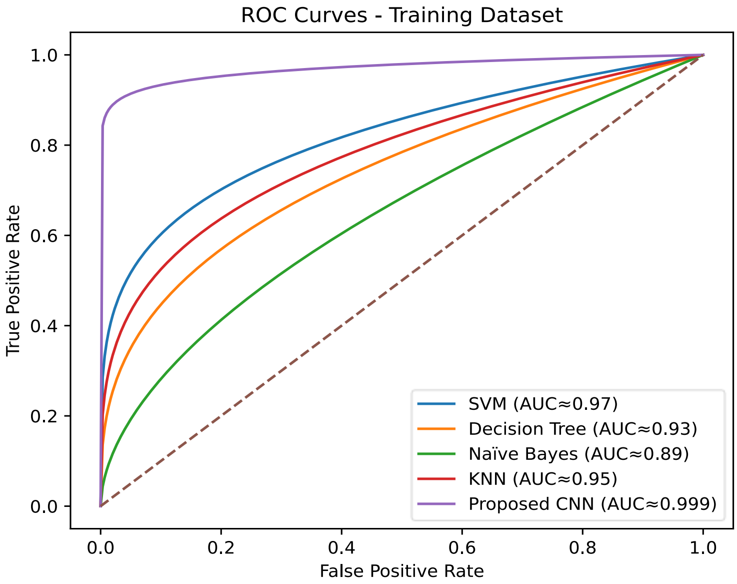


Figure 4. ROC-AUC plot for training data

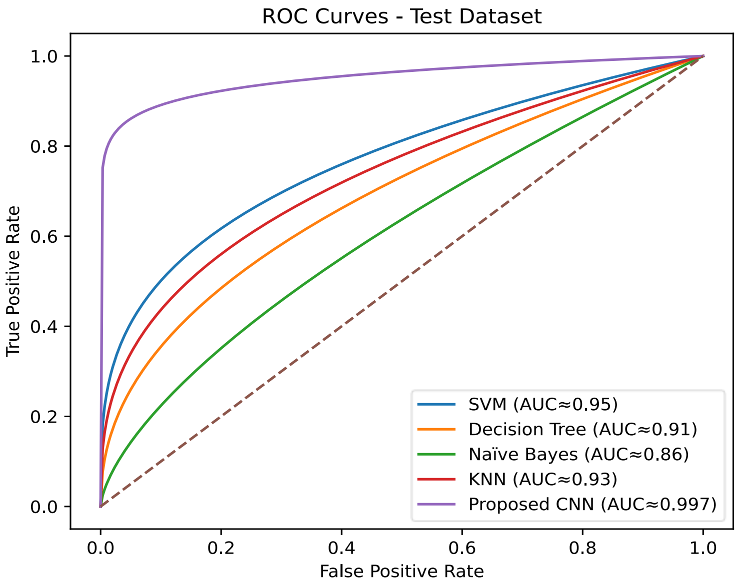


Figure 5. ROC-AUC plot for test data

Table 2. The five-fold cross-validation results

Fold	Accuracy (%)	F1-score (%)
Fold 1	99.6	99.4
Fold 2	99.7	99.5
Fold 3	99.8	99.6
Fold 4	99.6	99.4
Fold 5	99.7	99.5
Mean ± Std	99.68 ± 0.08	99.48 ± 0.08

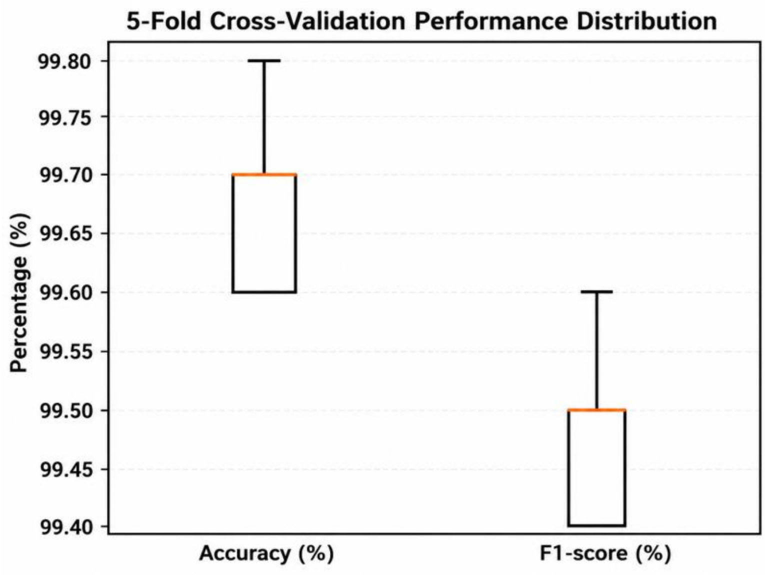


Figure 6. **Boxplot representation of Accuracy and F1-score obtained from 5-fold cross-validation of the proposed CNN model**

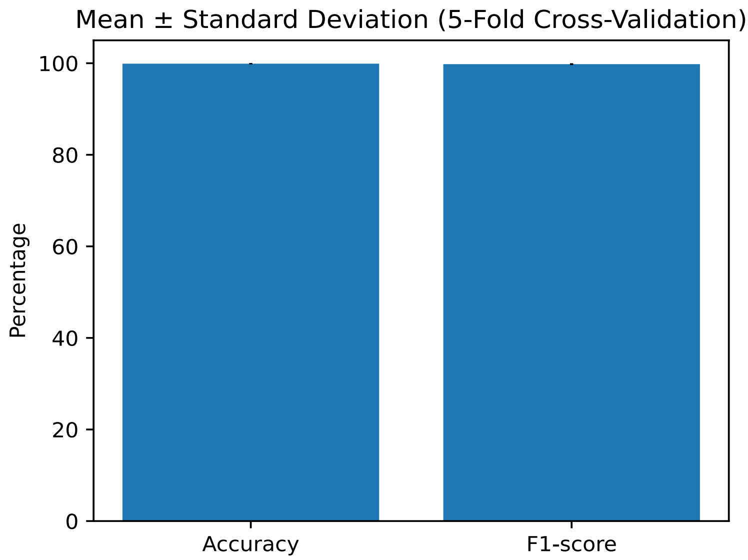


Figure 7. **Mean and standard deviation of Accuracy and F1-score obtained from 5-fold cross-validation**

In Figure 6, the narrow interquartile range indicates high stability and robustness across different data splits. In Figure 7, the small standard deviation demonstrates the strong generalization capability of the proposed CNN model.

The cross-validation analysis was not used for model selection or hyperparameter tuning, but solely to assess the robustness and generalizability of the final trained model.

6. Discussion

Patients with multiple sclerosis (MS), a specific patient population, are categorized into different stages of the disease. Determining the disease stage for each individual depends on their physician's assessment, which is achieved through a series of questions and physical examinations. In Khorasan Razavi Province, there is only one MS center and one MS fellowship-trained physician, both located at Qaem Hospital in Mashhad. Many MS patients visit the center daily to report their condition to the physician, who may prescribe new medications as needed. The MS center can become quite crowded on days when physicians visit, as patients come from all over Iran and often endure long waits, facing the challenges of the disease, to see their doctors. These challenges have motivated this research project, which aims to facilitate shorter patient-doctor meetings while providing more accurate monitoring and diagnosis of MS levels.

One precise method for identifying the level of MS is to use magnetic resonance imaging (MRI) and have a physician assess the patient's motor behavior. However, due to the high cost of MRI, many people cannot afford periodic expenses for these examinations. Therefore, this research investigates whether it is possible to identify the level of MS using standard questionnaires that encompass motor, behavioral, pharmaceutical, and other symptoms.

To achieve this without the physical presence of a physician, we required information from the patients. We collected this information from 470 patients through standardized questionnaires that had been previously tested and validated. These questionnaires covered all aspects of the disease, including mobility, urinary issues, sexual problems, sleep disturbances, quality of life, etc. We used the scores from these questionnaires as input data for this research.

To reduce the burden on patients with mobility issues, we aimed to reduce the number of questionnaires in the initial phase of this study. However, experimental results of a genetic algorithm showed that removing questionnaire scores, which serve as features in our context, led to a noticeable decrease in accuracy in identifying MS severity. Therefore, all questionnaire scores were used to identify MS severity in machine learning and deep learning models in the later stages of this research.

The proposed deep learning model achieved high accuracy in identifying MS severity in patients, as seen in the experimental results. Support vector machine (SVM) outperformed other machine learning algorithms, such as decision tree and naive Bayes and KNN, in terms of accuracy in identifying MS severity. High accuracy is essential in medical research, especially in this study, which aims to diagnose MS severity in patients. In deep learning, it is necessary to design an architecture tailored to the input data. The proposed architecture was successful in doing so.

It is possible to identify MS severity through questionnaires and patient responses. Still, this approach is challenging because it is difficult to assess the accuracy of patient responses without a physical examination. Therefore, in this research, we assumed that patients honestly responded to the questionnaires.

The second challenge in using questionnaires to identify MS severity is that they may not detect disease progression in patients with fewer visible symptoms. This can only be detected through MRI, which can visualize lesions in the brain. One solution is to require patients to undergo physical examination and imaging by a physician at specific intervals.

Although using questionnaires alone to identify MS severity can be challenging, they can be a valuable tool when patients complete them before visiting a specialist. Once patients visit a specialist, they have a documented history of their condition, and the physician can recommend medications based on more detailed physical examination and imaging, such as MRI. Therefore, questionnaires can expedite the examination and prescription process, reducing patient wait times.

Additionally, because patient medical histories can be recorded in a database, physicians can compare patients' current status to their previous records during each visit. In critical cases, such as excessive sleep disturbances, severe depression, or suicidal ideation, physicians can take necessary preventive measures more rapidly.

Figure 8 shows the response of a patient who improved from MS severity level 4 to level 3. The diagrams in Figure 8 show the patient's primary response (left bar- green) and secondary response (right bar- blue) after 4 months. Each diagram shows the difference in the patient's scores for each questionnaire.

It is well-known that classical machine learning algorithms often struggle with certain data distributions, and their performance is highly sensitive to parameter tuning. They are more prone to underfitting or overfitting when applied to complex or high-dimensional data. In contrast, deep learning models offer greater flexibility through architecture design, layer depth, and hyperparameter optimization, allowing them to capture intricate patterns in the data better.

In our study, the proposed deep learning model was carefully designed and tuned through experimental evaluation on the collected dataset. The high accuracy achieved on the test set resulted from selecting an appropriate architecture tailored to the data's specific characteristics.

That said, we fully acknowledge the potential risk of overfitting, especially when the model achieves near-perfect accuracy. To mitigate this, we employed strategies such as strict separation of training and test data, dropout regularization, a controlled number of training epochs, and a relatively small and specialized model architecture. We believe these measures helped ensure the model's generalization. However, we also note that as the dataset grows in future work, further hyperparameter tuning may be necessary to maintain optimal performance—an expected and common practice in deep learning model development.

Despite the promising results, this study has several limitations that should be acknowledged. Although 470 MS patients initially completed the questionnaires, only 152 responses with complete and reliable data were selected for model training and evaluation. The deep learning model was built using 20 input features derived from these patient-reported outcomes. Moreover, all data were collected from a single medical center (Qaem Hospital), which may limit the generalizability of the findings to broader or more diverse MS populations. Therefore, while the model demonstrated excellent performance within this dataset, further validation on larger, multi-center, and demographically varied cohorts is essential to confirm its robustness and broader clinical applicability.

6.1. Limitations

This study has several limitations. First, the dataset size is relatively small and was collected from a single center, which may limit generalizability. Second, external validation on independent cohorts was not performed. Future work will address these limitations by incorporating larger, multi-center datasets and conducting external and prospective validation studies.

The proposed framework is intended to support clinical decision-making rather than replace clinician judgment. Its role is to assist in identifying patients who may require closer monitoring or further evaluation. Prospective clinical studies are needed to assess its real-world impact on clinical workflows and patient outcomes.

7. Conclusion

This study proposed a deep learning–based framework for assessing the severity of multiple sclerosis (MS) using patient-reported outcomes. By integrating responses from 16 standardized questionnaires covering motor function, psychological state, fatigue, and behavioral factors, the proposed CNN model demonstrated consistently high classification performance, achieving accuracy exceeding 99.5% across experimental evaluations.

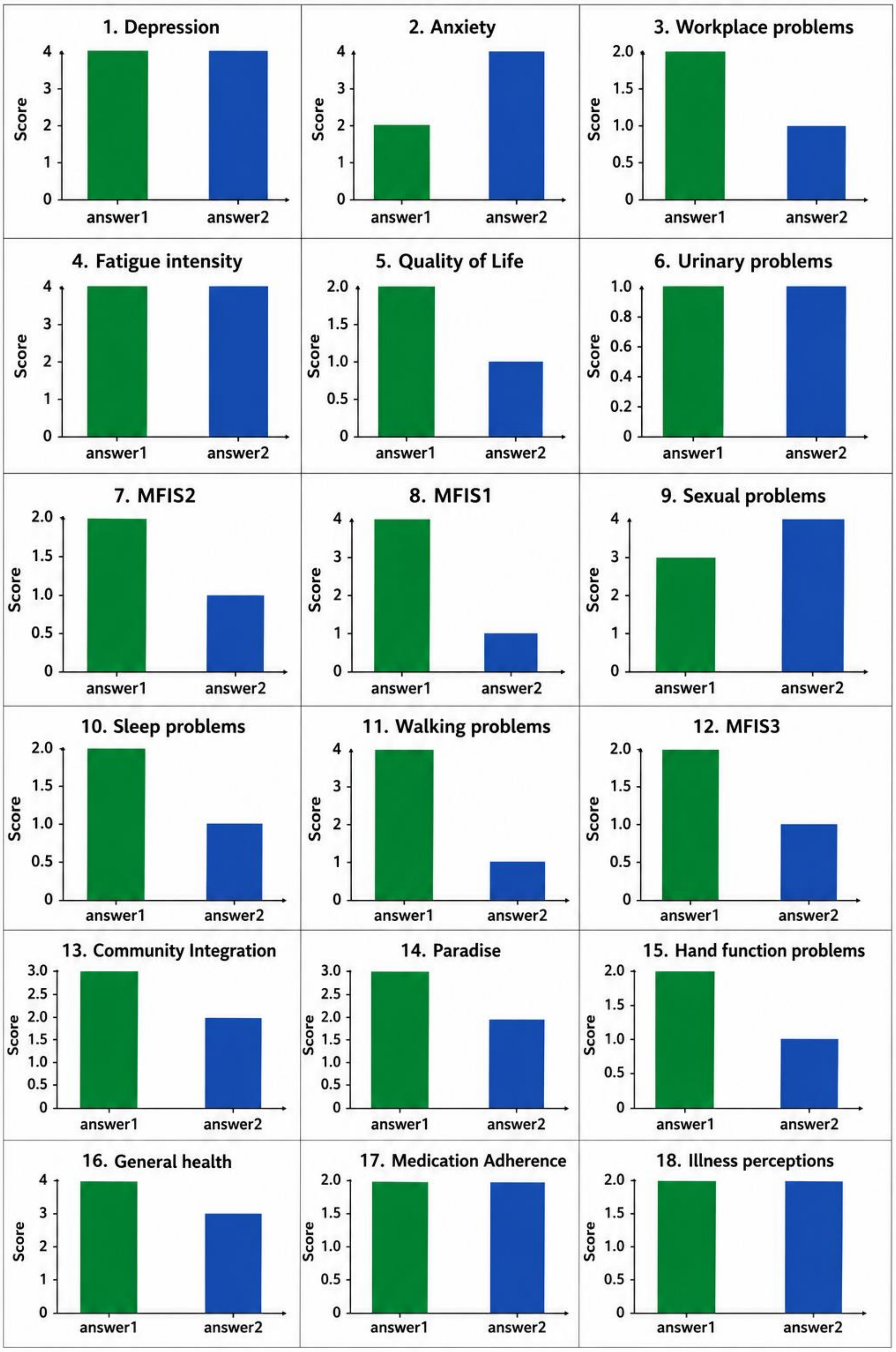


Figure 8. Comparison of questionnaire scores of MS patients who have improved from level 4 to level 3

Comparative analyses showed that the proposed approach outperformed conventional machine learning methods, including decision trees, naïve Bayes, K-nearest neighbors (KNN), and support vector machines (SVM), highlighting the capability of deep learning models to effectively capture complex, multidimensional symptom patterns derived from questionnaire-based data.

Rather than replacing clinical judgment, the proposed framework is intended to function as a decision-support tool that may assist clinicians in preliminary severity assessment and patient stratification before in-person consultation. Feedback from expert neurologists indicated that such a system could support pre-consultation triage and longitudinal symptom tracking, particularly in settings where frequent neurological evaluations are challenging.

Despite the promising retrospective results, this study is subject to several limitations, including a relatively small, single-center dataset and the absence of external validation. Future work will focus on validating the framework on larger, multi-center cohorts, incorporating standardized clinical scales such as EDSS for comparative analysis, and conducting prospective studies to evaluate its real-world clinical utility and impact on patient outcomes.

8. Statements & Declarations

8.1. Author Contributions

All authors have read and agreed to the published version of the manuscript.

8.2. Data Availability Statement

Data available on request due to restrictions eg privacy or ethical: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions concerning the confidentiality of patient information.

8.3. Funding

This research received no external funding.

8.4. Acknowledgments

The authors have no acknowledgments to declare.

8.5. Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The authors used ChatGPT (OpenAI) to assist in refining the language, improving clarity, and structuring responses to reviewers' comments. The AI tool was not used for data analysis, model development, or result generation.

8.6. Ethics Approval

Not applicable.

8.7. Informed Consent

Informed consent was obtained from all participants.

8.8. Consent for Publication

Not applicable.

8.9. Conflicts of Interest

The authors declare no conflict of interest.

9. References

[1] J. Oh, A. Vidal-Jordana, and X. Montalban, "Multiple sclerosis: Clinical aspects," Curr. Opin. Neurol., vol. 31, no. 6, pp. 752–759, 2018, doi:10.1097/WCO.0000000000000622.

[2] S. Rezaei, A. Saberi, H. Hatamian, A. Ghayeghran, Z. Khaksari, and F. Mollahosein, "Adaptation and validation of the Hamburg Quality of Life Questionnaire in Multiple Sclerosis (HAQUAMS) for use in Iranian patients," Acta Neurologica Taiwanica, vol. 31, no. 1, pp. 24–35, Jan. 2022.

[3] N. Asmarian et al., "Multiple sclerosis incidence rate in southern Iran: a Bayesian epidemiological study," BMC Neurol., vol. 21, no. 1, p. 309, 2021, doi:10.1186/s12883-021-02342-1.

[4] S. DARVISHI, O. HAMIDI, and J. POOROLAJAL, "Prediction of Multiple sclerosis disease using machine learning classifiers: A comparative study," J. Prev. Med. Hyg., vol. 62, no. 1, pp. E192–E199, 2021, doi:10.15167/2421-4248/jpmh2021.62.1.1651.

[5] N. Fattahi et al., "Burden of multiple sclerosis in Iran from 1990 to 2017," BMC Neurol., vol. 21, no. 1, p. 400, 2021, doi:10.1186/s12883-021-02431-1.

[6] M. Filippi et al., "Prediction of a multiple sclerosis diagnosis in patients with clinically isolated syndrome using the 2016 MAGNIMS and 2010 McDonald criteria: a retrospective study," Lancet Neurol., vol. 17, no. 2, pp. 133–142, 2018, doi:10.1016/S1474-4422(17)30469-6.

[7] S. Afshar et al., "Validity and reliability of Persian version of the Arm Function in Multiple Sclerosis Questionnaire," Br. J. Occup. Ther., vol. 85, no. 2, pp. 130–136, 2022, doi:10.1177/03080226211008710.

[8] C. C. Lin, L. Muthukumar, E. L. Reynolds, C. E. Hill, G. J. Esper, and B. C. Callaghan, "Wait Time to See a Neurologist After Referral Among Medicare Participants," Neurology, vol. 104, no. 3, p. 210217, 2025, doi:10.1212/WNL.0000000000210217.

[9] F. Khodaie et al., "Time interval between the onset of symptoms and diagnosis of multiple sclerosis and the influential factors: A national registry-based study," Clin. Neurol. Neurosurg., vol. 239, p. 108221, 2024, doi:10.1016/j.clineuro.2024.108221.

[10] N. Sadeghi et al., "Innovating Care in Multiple Sclerosis: Feasibility of Synchronous Internet-Based Teleconsultation for Longitudinal Clinical Monitoring," J. Pers. Med., vol. 12, no. 3, p. 433, 2022, doi:10.3390/jpm12030433.

[11] K. Zhang, J. A. Lincoln, X. Jiang, E. V. Bernstam, and S. Shams, "Predicting multiple sclerosis severity with multimodal deep neural networks," BMC Med. Inform. Decis. Mak., vol. 23, no. 1, 2023, doi:10.1186/s12911-023-02354-6.

[12] S. Roy et al., "Performance of machine learning models for predicting high-severity symptoms in multiple sclerosis," Sci. Rep., vol. 15, no. 1, p. 18209, 2025, doi:10.1038/s41598-024-63888-x.

[13] P. Schwab and W. Karlen, "A Deep Learning Approach to Diagnosing Multiple Sclerosis from Smartphone Data," IEEE J. Biomed. Heal. Informatics, vol. 25, no. 4, pp. 1284–1291, 2021, doi:10.1109/JBHI.2020.3021143.

[14] Hoeijmakers et al., "Disease severity classification using passively collected smartphone-based keystroke dynamics within multiple sclerosis," Sci. Rep., vol. 13, no. 1, 2023, doi:10.1038/s41598-023-28990-6.

[15] S. Gustavsen et al., "The association of selected multiple sclerosis symptoms with disability and quality of life: a large Danish self-report survey," BMC Neurol., vol. 21, no. 1, p. 317, 2021, doi:10.1186/s12883-021-02344-z.

[16] M. Vázquez-Marrufo, E. Sarrias-Arrabal, M. García-Torres, R. Martín-Clemente, and G. Izquierdo, "A systematic review of the application of machine-learning algorithms in multiple sclerosis," 2023. doi:10.1016/j.nrleng.2020.10.013.

[17] O. Mirmosayyeb, V. Shaygannejad, S. Najdaghi, and M. Ghajarzadeh, "Psychometric Properties of the Persian Version of the Multiple Sclerosis Work Difficulties Questionnaire (MSWDQ)," Int. J. Prev. Med., vol. 13, no. 1, p. 94, 2022, doi:10.4103/ijpvm.ijpvm_393_20.

[18] S. Kaviani and V. Sharifi, "Ebrahimkhani: Reliability and validity of the hospital anxiety and depression scale (hads): Iranian depressed and anxious patients," J. Fac. Med. Tehran Univ. Med. Sci., vol. 67, no. 5, pp. 379–385. (In Persian).

[19] M. Azadvari, S. Z. E. Razavi, M. Shahrooei, A. N. Moghadasi, A. Azimi, and H. R. Farhadi-Shabestari, "Bladder dysfunction in Iranian patients with multiple sclerosis," J. Multidiscip. Healthc., vol. 13, pp. 345–349, 2020, doi:10.2147/JMDH.S244697.

[20] P. V. Rabins and B. R. Brooks, "Emotional disturbance in multiple sclerosis patients: Validity of the General Health Questionnaire (GHQ)," Psychol. Med., vol. 11, no. 2, pp. 425–427, 1981, doi:10.1017/S0033291700052259.

[21] F. Gharibi et al., "Quality of Life and Its Relative Factors Among Patients With Multiple Sclerosis: A Cross-sectional Study in Northwest Iran," J. Res. Heal., vol. 13, no. 4, pp. 263–272, 2023, doi:10.32598/JRH.13.4.2197.1.

[22] M. Heidari, M. Akbarfahimi, M. Salehi, and S. M. Nabavi, "Validity and reliability of the Persian-version of fatigue impact scale in multiple sclerosis patients in Iran," Koomesh, vol. 15, no. 3, pp. 295–301, 2014, [Online]. Available: https://brieflands.com/journals/koomesh/articles/152647

[23] M. Azadvari, A. N. Moghadasi, S. Navardi, S. Z. E. Razavi, and S. H. Gashti, "Relationship Between Bowel/bladder, and Sexual Dysfunction With Health-Related Quality of Life in Women With Multiple Sclerosis (MS).," Dec. 2020, doi:10.21203/rs.3.rs-128355/v1.

[24] S. M. Nabavi et al., "Prevalence of Sexual Dysfunction and Related Risk Factors in Men with Multiple Sclerosis in Iran: A Multicenter Study," Neurol. Ther., vol. 10, no. 2, pp. 711–726, 2021, doi: 10.1007/s40120-021-00257-0.

[25] M. Ghajarzadeh, R. Jalilian, G. Eskandari, M. Ali Sahraian, and A. Reza Azimi, "Validity and reliability of Persian version of Modified Fatigue Impact Scale (MFIS) questionnaire in Iranian patients with multiple sclerosis," Disabil. Rehabil., vol. 35, no. 18, pp. 1509–1512, 2013, doi:10.3109/09638288.2012.742575.

[26] S. Rahimibarghani, M. Azadvari, S. Z. Emami-Razavi, M. H. Harirchian, S. Rahimi-Dehgolan, and H. R. Fateh, "Effects of Nonconsecutive Sessions of Transcranial Direct Current Stimulation and Stationary Cycling on Walking Capacity in Individuals With Multiple Sclerosis," Int. J. MS Care, vol. 24, no. 5, pp. 202–208, 2022, doi:10.7224/1537-2073.2021-004.

[27] N. Razaziyan, F. Najafi, P. Mahdavi, and A. Aghaei, "Prevalence of sleep disorders in patients with multiple sclerosis," J. Maz. Univ. Med. Sci., vol. 23, no. 110, pp. 219–224, 2014. (In Persian).

[28] Azimi et al., "Psychometric Properties of the Persian Version of the PARADISE-24 Questionnaire," Int. J. Prev. Med., vol. 12, no. 1, p. 50, 2021, doi:10.4103/ijpvm.IJPVM_300_19.

[29] H. Baharlouei, M. Salavati, B. Akhbari, Z. Mosallanezhad, M. Mazaheri, and H. Negahban, "Cross-cultural validation of the Falls Efficacy Scale International (FES-I) using self-report and interview-based questionnaires among Persian-speaking elderly adults," Arch. Gerontol. Geriatr., vol. 57, no. 3, pp. 339–344, 2013, doi:10.1016/j.archger.2013.06.005.

[30] Y. Moharamzad et al., "Validation of the Persian Version of the 8-Item Morisky Medication Adherence Scale (MMAS-8) in Iranian Hypertensive Patients," Glob. J. Health Sci., vol. 7, no. 4, pp. 173–183, 2015, doi:10.5539/gjhs.v7n4p173.

[31] H. Negahban, P. Fattahizadeh, R. Ghasemzadeh, R. Salehi, N. Majdinasab, and M. Mazaheri, "The Persian version of community integration questionnaire in persons with multiple sclerosis: Translation, reliability, validity, and factor analysis," Disabil. Rehabil., vol. 35, no. 17, pp. 1453–1459, 2013, doi:10.3109/09638288.2012.741653.

[32] N. Masaeli, R. Bagherian, G. Kheirabadi, A. Khedri, and B. Mahaki, "Psychometric properties of the Persian version of the Brief Illness Perception Questionnaire in chronic diseases," Journal of Research in Rehabilitation Sciences, vol. 14, no. 6, pp. 332–337, 2019, doi:10.22122/jrrs.v14i6.3266.

[33] Marzullo et al., "Classification of multiple sclerosis clinical profiles via graph convolutional neural networks," Front. Neurosci., vol. 13, no. JUN, p. 594, 2019, doi:10.3389/fnins.2019.00594.

[34] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights Imaging, vol. 9, no. 4, pp. 611–629, 2018, doi:10.1007/s13244-018-0639-9.