



e-ISSN:2582-7219



# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 11, November 2024



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

Impact Factor: 7.521



6381 907 438



6381 907 438



ijmrset@gmail.com



www.ijmrset.com



## International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# Artificial Intelligence ML KNN Subspace Parameters Performance in Classification of Sclerosis

Loida F. Hermosura, CpE, MSIT

College of Information Technology, Northeastern College, Santiago City, Isabela, Philippines

**ABSTRACT:** This study investigated the efficacy of a Subspace KNN ensemble model for the classification of multiple sclerosis (MS), specifically differentiating between clinically definite multiple sclerosis (CDMS) and non-CDMS cases. Using a dataset of 547 patient records with 19 clinical, demographic, and diagnostic features, the model demonstrated promising performance with a high validation accuracy of 97.3%, rapid prediction speed, and efficient training time. However, analysis of the confusion matrix revealed the presence of false positives and the absence of a false negative value, highlighting the need for a comprehensive evaluation approach that considers various performance metrics beyond accuracy. Further research is recommended to optimize hyperparameters, incorporate cost-sensitive learning, analyze feature importance, and evaluate the model on diverse datasets to enhance its performance and clinical utility. Employing Explainable AI techniques can also improve transparency and trust in the model's predictions. This study contributes to the development of accurate and reliable tools for MS diagnosis and management, ultimately aiming to improve patient care and outcomes.

**KEYWORDS:** Sclerosis, Machine Learning, Subspace, KNN, Confusion Matrix.

## I.INTRODUCTION

The most common neurological impairment, multiple sclerosis (MS), is an autoimmune-mediated condition that affects the central nervous system (CNS) and frequently causes neurological issues in young adults in addition to significant physical or cognitive incapacitation [1]. The disease first manifests in young adults, with those in their 20s and 30s being the most susceptible [2-3]. Every year, MS affects over 2.5 million people globally [4]. Our incomplete knowledge of the molecular mechanisms behind multiple sclerosis (MS), the absence of reliable prognostic or predictive biomarkers, and the clinical variability among patients all impede the development of individualized healthcare for MS patients [5-7]. Early diagnosis of diseases like MS is vital for improving patient survival rates by ensuring a higher proportion of cases are identified at the initial stages. Traditional classification techniques such as logistic regression have been utilized extensively in medical research to distinguish between cases and controls. Although these models provide clear and interpretable results, they often fail to capture intricate relationships between variables, limiting their predictive power. Consequently, there is an increasing demand for more advanced machine learning models that offer superior accuracy and minimal prediction error. Emerging techniques like deep learning, decision trees, and ensemble methods are increasingly preferred in the medical field due to their ability to model complex, non-linear relationships. These models can automatically detect patterns in large datasets, making them ideal for applications in medical diagnosis, where subtle differences in patient data can be crucial for early detection. Moreover, the integration of advanced computational methods with real-time data collection, such as from wearable devices and medical imaging, further enhances the precision of these systems. As a result, modern medical diagnostic tools are becoming more sophisticated, adaptive, and reliable, significantly improving the chances of early disease detection and patient outcomes. Certain biomarkers linked to multiple sclerosis (MS) have been demonstrated to have a strong predictive value of a more severe course of the disease. These include the presence of oligoclonal IgM bands [8,9], neurofilaments light [10], or chitinase-3 [11] in the serum and cerebrospinal fluid (CSF).

The labels benign and malignant, which are markers of the disease severity over time rather than a regular pattern of classification, are also commonly used to analyze the course of MS. While the malignant form of MS is characterized by numerous incapacitating episodes and incomplete recovery, which leads to a rapid progression of disability, the benign form is typically characterized by few relapses and reduced/absence of disability after 20 years of evolution. Because different specialists frequently utilize different definitions using the Expanded Disability Status Scale (EDSS),



## International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the indicators criteria are not exact [12-13]. Machine Learning Nowadays are used in different studies such as in agriculture [14-16] and medicine [17-18].

### II. RELATED LITERATURE

A review of previous research on MS diagnosis using AI techniques has been conducted by a number of papers, including [19], which examined the majority of earlier studies that employed DL algorithms for the automated diagnosis of MS using MRI data. In addition to outlining the current difficulties and potential avenues for future research, they talked about the most popular preprocessing methods.

Finding the most effective approaches and strategies for MS diagnosis was the goal of Arani et al. [20]. The writers evaluated the effectiveness of those approaches to suggest the best one. They discovered that the most popular approaches for diagnosing multiple sclerosis (MS) are rule-based, fuzzy logic (FL), and artificial neural networks (ANN). They also noted the shortcomings of each of these approaches and suggested combining them to overcome their shortcomings and increase the diagnostic systems' accuracy.

The following are just a few of the many ways that DL and ML have helped clinicians throughout the history of medicine: first, by identifying individuals who are at risk for the disease and warning them to avoid triggers; second, by accurately and early diagnosing the disease, which leads to the use of therapeutic agents that are known to delay the prognosis of the disease and thereby improve the quality of life of those patients; third, by predicting the progression of the disease from one mild type to another based on the analysis of various blood, cerebrospinal fluid (CSF), and radiological markers; and fourth, by forecasting the effectiveness of specific medications in preventing the disease's progression and treatment monitoring.

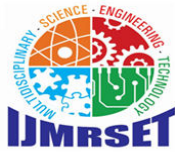
Sarbaz et al. [21] created a decision support system (DSS) that uses a straightforward, noninvasive approach to identify MS patients who depend on balance disorders. Twenty healthy controls and fourteen MS patients were enrolled in that study. Every participant had a marking placed between their eyebrows on their forehead. After then, participants stood in front of a black background for three minutes while being videotaped. An image processing technique was used to examine and analyze the relocation of these markers. An ANN with a "tan-sigmoid" transfer function was employed. Finding the characteristics that demonstrated a substantial difference between the MS patients and the healthy controls was essential to feature extraction. 92.35% accuracy was attained by the ANN.

Recently, MS was diagnosed using ensemble learning, a classification technique, with a noteworthy 94.91% accuracy rate. This study used DT-based ensemble learning for classification and 18 distinct GLCM features for feature extraction. AdaBoost, LogitBoost, and LPBoost were the three boosting techniques used to categorize MR pictures of healthy and diseased brains. From preprocessing to classification across a dataset of 293 images, the study's methodology stands out for its thorough approach and excellent performance metrics, proving its superiority over conventional neural network and wavelet transform techniques. [22]

### III. METHODOLOGY

In this study, an open-source dataset from Kaggle was utilized, consisting of 547 samples to explore the relationships between clinical variables and the diagnosis of clinically definite multiple sclerosis (CDMS). The dataset contains patient data including demographic information, medical history, and various diagnostic test results. Each record provides detailed features such as age, schooling duration, gender, breastfeeding history, and test results from neurophysiological assessments like visual evoked potentials (VEP), brainstem auditory evoked potentials (BAEP), and somatosensory evoked potentials (SSEP).

The dataset includes key diagnostic variables such as the age of the patient, measured in years, along with the time spent in formal education, represented by the number of years of schooling. Gender is coded as 1 for male and 2 for female, while breastfeeding history is categorized into yes, no, or unknown, with corresponding values of 1, 2, and 3. In



## International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

addition, the dataset includes information on whether the patient had a history of varicella, also known as chickenpox, which is recorded as positive, negative, or unknown.

The dataset captures initial symptoms, such as whether they were visual, sensory, motor, or a combination thereof, along with neurophysiological test results like brainstem auditory evoked potentials (BAEP), visual evoked potentials (VEP), and somatosensory evoked potentials (SSEP) from the upper and lower limbs. Other important diagnostic results include the presence of oligoclonal bands in cerebrospinal fluid, periventricular MRI results, and findings from cortical, infratentorial, and spinal cord MRIs. Each sample is classified into one of two groups, representing patients with clinically definite multiple sclerosis (CDMS) or non-CDMS cases.

The primary goal of this study is to apply machine learning techniques to predict the likelihood of CDMS based on these features, using decision tree-based algorithms to assess their diagnostic accuracy.

Gender	Age	Schooling	Breastfeec	Varicella	Initial_Sym	Mono_or_F	Oligoclon	LLSSEP	ULSSEP	VEP	BAEP	Periventric	Cortical_M	Infratentor	Spinal_Cor	Initial_EDS	Final_EDS	group	
1	34	20	1	1	2	1	0	1	1	0	0	0	1	0	1	1	1	1	
1	61	25	3	2	10	2	1	1	0	1	0	0	0	0	1	2	2	1	
1	22	20	3	1	3	1	1	0	0	0	0	0	1	0	0	1	1	1	
2	41	15	1	1	7	2	1	0	1	1	1	0	1	1	0	0	1	1	
2	34	20	2	1	6	2	0	1	0	0	0	1	0	0	0	1	1	1	
1	29	22	1	1	6	2	0	1	0	0	0	1	0	1	0	1	1	1	
2	53	20	1	1	14	2	0	1	0	1	0	1	1	0	1	1	1	1	
2	24	15	1	1	14	2	0	1	1	0	0	1	1	1	1	2	2	1	
1	36	15	1	1	8	2	0	1	1	1	1	0	1	0	0	1	1	1	
2	28	20	1	1	8	2	0	0	0	0	0	1	0	1	0	1	1	1	
2	60	12	3	2	15	2	0	1	0	0	0	1	0	0	1	1	1	1	
2	25	20	1	1	5	2	0	1	0	1	1	0	0	1	0	1	1	1	
1	34	12	1	1	11	2	0	1	1	1	1	0	1	1	1	0	2	2	1
1	36	0	1	1	13	2	1	0	0	0	0	1	0	1	0	1	1	1	1
2	29	15	1	1	1	1	0	1	1	1	1	0	1	1	1	1	2	2	1
1	29	12	1	2	5	2	0	1	1	1	1	0	0	1	0	0	1	1	1
2	29	20	1	1	15	2	0	1	1	1	1	0	1	0	1	1	3	3	1
2	24	20	1	1	8	2	1	0	0	1	1	0	0	1	1	0	1	1	1
2	51	9	1	2	10	2	0	1	1	1	1	0	1	0	0	1	2	3	1
2	36	20	1	1	8	2	1	0	0	0	0	1	1	1	1	0	1	1	1
2	32	15	3	1	11	2	1	0	0	0	0	1	0	0	0	1	1	1	1
2	30	15	3	2	15	2	1	0	0	0	0	1	0	0	1	2	2	1	1
1	50	12	1	2	13	2	1	1	1	1	1	0	1	1	1	1	2	2	1
1	38	15	1	2	11	2	1	0	1	1	1	0	1	1	0	0	1	1	1
1	22	20	3	2	5	2	0	1	0	1	1	0	1	1	1	1	1	1	1

Figure 1. MS Dataset

In figure 1 shows the dataset what were used in classification of multiple sclerosis. The dataset used in this study, as displayed in the Figure 1, consists of 547 samples with 19 columns representing various clinical, demographic, and diagnostic features. Each row corresponds to a single patient record, where features include gender, age, schooling duration, breastfeeding history, varicella (chickenpox) history, and the nature of the initial symptoms experienced by the patient. The dataset also contains neurophysiological test results, such as the presence of oligoclonal bands, upper and lower limb somatosensory evoked potentials (LLSSEP and ULSSEP), visual evoked potentials (VEP), and brainstem auditory evoked potentials (BAEP).

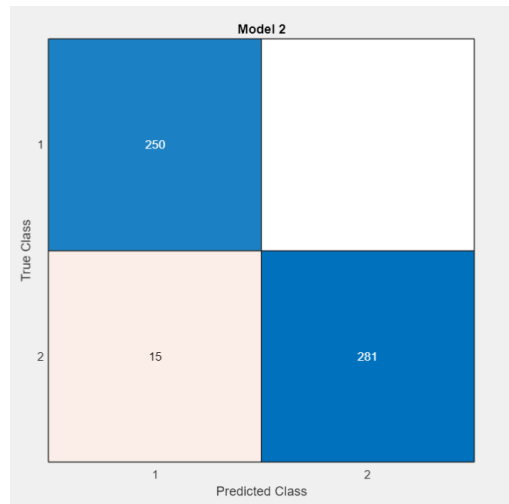
Further columns record MRI results, such as periventricular, cortical, infratentorial, and spinal cord MRI findings. The final columns represent the Expanded Disability Status Scale (EDSS) scores at the initial and final stages, which measure the level of disability in multiple sclerosis patients, and a grouping variable that categorizes patients into two groups: clinically definite multiple sclerosis (CDMS) and non-CDMS cases. These diverse features provide a rich dataset for analyzing the likelihood of a CDMS diagnosis using machine learning techniques.



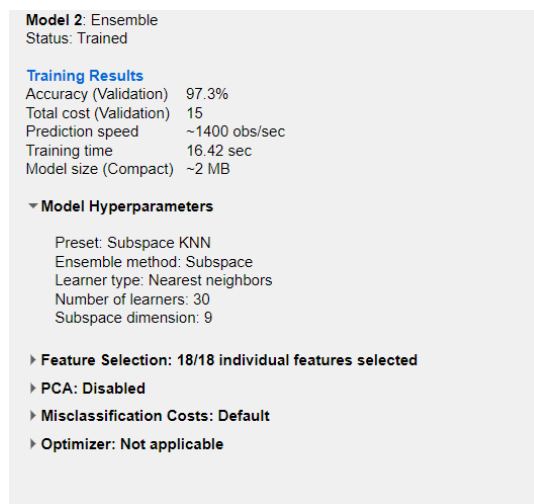
## International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

### IV. RESULTS AND DISCUSSION



(a). Model’s Confusion Matrix using Subspace KNN Parameters



(b.) Results of Subspace KNN and Parameters

The matrix displays a strong initial performance with a high number of true negatives (250) and true positives (281). This suggests that the model is generally accurate in identifying both individuals who do not have sclerosis and those who do. However, a critical observation is the presence of 15 false positives, indicating that the model incorrectly classified 15 individuals as having sclerosis when they actually did not. This raises concerns about the potential consequences of such misdiagnoses, including unnecessary anxiety, further testing, and potentially harmful treatments. Crucially, the matrix lacks the value for false negatives, representing the number of individuals incorrectly classified as not having sclerosis when they do. This missing information hinders a complete assessment of the model's performance. False negatives in this context are particularly worrisome due to the potential for delayed diagnosis and treatment of MS, which can lead to disease progression and increased disability. The training process of Model 2 yielded promising results. A high validation accuracy of 97.3% suggests that the model generalizes well to unseen data and is capable of accurately distinguishing between individuals with and without sclerosis. This is further supported by a relatively low total cost (validation) of 15, although the specific meaning of this metric requires further clarification.



## International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The model exhibits impressive efficiency, with a prediction speed of approximately 1400 observations per second. This rapid processing capability is crucial for real-world applications where timely diagnosis is essential. Furthermore, the model boasts a compact size of approximately 2 MB, making it suitable for deployment across various platforms and devices.

The training time of 16.42 seconds is notably short, indicating efficient learning despite the complexity of the ensemble approach. This efficiency can be attributed to the chosen hyperparameters and the inherent properties of the Subspace KNN algorithm.

### V.CONCLUSION

This study investigated the application of a Subspace KNN ensemble model for the classification of multiple sclerosis (MS), specifically focusing on the differentiation between clinically definite multiple sclerosis (CDMS) and non-CDMS cases. Utilizing a dataset comprising 547 patient records with 19 clinical, demographic, and diagnostic features, the model demonstrated promising results in terms of accuracy, efficiency, and training time.

The Subspace KNN ensemble, with its unique approach of combining multiple k-nearest neighbors' learners within specific subspaces of the data, achieved a high validation accuracy of 97.3%. This indicates the model's ability to effectively learn from the provided data and generalize well to unseen cases. The model's efficiency, with a prediction speed of ~1400 observations per second and a compact size of ~2 MB, further highlights its potential for real-world clinical applications where timely and accessible diagnosis is crucial. However, the analysis of the model's performance through a confusion matrix revealed critical aspects that require further consideration. While the model exhibited a high number of true positives and true negatives, the presence of 15 false positives raises concerns about the potential for misdiagnosis and its associated consequences, such as unnecessary anxiety, further testing, and potentially harmful treatments. More importantly, the absence of the false negative value in the confusion matrix hinders a complete assessment of the model's sensitivity, i.e., its ability to correctly identify individuals with CDMS. False negatives in this context are particularly significant due to the potential for delayed diagnosis and treatment, which can lead to disease progression and increased disability. The discrepancy between the high validation accuracy and the presence of false positives underscores the importance of a multifaceted evaluation approach that goes beyond relying solely on accuracy as a performance metric. A comprehensive assessment necessitates considering other metrics, such as sensitivity, specificity, and precision, to gain a holistic understanding of the model's strengths and weaknesses. The selection and fine-tuning of hyperparameters, including the number of learners, subspace dimension, and misclassification costs, play a crucial role in optimizing the model's performance and mitigating the risk of misclassifications. Future research should focus on systematically exploring the hyperparameter space to identify optimal settings that minimize both false positives and false negatives. Furthermore, incorporating cost-sensitive learning, where different costs are assigned to false positives and false negatives, can guide the model to prioritize the correct classification of individuals with CDMS, even if it leads to a slight increase in false positives. Analyzing feature importance can provide valuable insights into the model's decision-making process and potentially lead to the identification of new, more informative features. Evaluating the model on diverse datasets, including those with varying demographics and disease subtypes, can assess its generalizability and robustness across different populations. Finally, employing Explainable AI (XAI) techniques can enhance transparency and trust in the model's predictions by providing insights into its decision-making process. This is particularly important in medical applications where understanding the rationale behind a diagnosis is crucial for both clinicians and patients. In conclusion, the Subspace KNN ensemble model presents a promising approach for the detection of CDMS. However, addressing the challenges related to false positives and false negatives, through further research and refinement of the model, is essential to ensure its clinical utility and improve patient outcomes. By combining advanced machine learning techniques with a comprehensive evaluation strategy and a focus on explainability, researchers can pave the way for the development of accurate, reliable, and trustworthy tools for MS diagnosis and management.

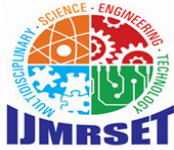


## International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

### REFERENCES

1. M. Zorzon et al., "Risk factors of multiple sclerosis: a case-control study," *Neurological Sciences*, vol. 24, no. 4, pp. 242–247, Nov. 2003, doi: 10.1007/s10072-003-0147-6.
2. Etemadifar M, Sajjadi S, Nasr Z, Firoozeei TS, Abtahi SH, Akbari M, Fereidan-Esfahani M. Epidemiology of multiple sclerosis in Iran: a systematic review. *Eur Neurol* 2013;70:356-63. <https://doi.org/10.1159/000355140>
3. Harbo HF, Gold R, Tintoré M. Sex and gender issues in multiple sclerosis. *Ther Adv Neurol Diso* 2013;6:237-48. <https://doi.org/10.1177/1756285613488434>
4. Di Cara M, Lo Buono V, Corallo F, Cannistraci C, Rifichi C, Sessa E, D'Aleo G, Bramanti P, Marino S. Body image in multiple sclerosis patients: a descriptive review. *Neurol Sci* 2019;40:923-8. <https://doi.org/10.1007/s10072-019-3722-1>
5. E. Kotelnikova et al., "Dynamics and heterogeneity of brain damage in multiple sclerosis," *PLoS Computational Biology*, vol. 13, no. 10, p. e1005757, Oct. 2017, doi: 10.1371/journal.pcbi.1005757.
6. I. Pulido-Valdeolivas, I. Zubizarreta, E. H. Martinez-Lapiscina, and P. Villoslada, "Precision medicine for multiple sclerosis: an update of the available biomarkers and their use in therapeutic decision making," *Expert Review of Precision Medicine and Drug Development*, vol. 2, no. 6, pp. 345–361, Oct. 2017, doi: 10.1080/23808993.2017.1393315.
7. P. Villoslada, "Personalized medicine for multiple sclerosis: How to integrate neurofilament light chain levels in the decision?," *Multiple Sclerosis Journal*, vol. 27, no. 13, pp. 1967–1969, Oct. 2021, doi: 10.1177/13524585211049552.
8. Villar LM, Casanova B, Ouamara N et al (2014) Immunoglobulin M oligoclonal bands: biomarker of targetable inflammation in primary progressive multiple sclerosis. *Ann Neurol* 76:231–240
9. A. Huss et al., "Intrathecal immunoglobulin M production: A promising high-risk marker in clinically isolated syndrome patients," *Annals of Neurology*, vol. 83, no. 5, pp. 1032–1036, Apr. 2018, doi: 10.1002/ana.25237.
10. J. Kuhle et al., "Blood neurofilament light chain as a biomarker of MS disease activity and treatment response," *Neurology*, vol. 92, no. 10, Feb. 2019, doi: 10.1212/wnl.0000000000007032.
11. S. Brune et al., "Serum neurofilament light chain concentration predicts disease worsening in multiple sclerosis," *Multiple Sclerosis Journal*, vol. 28, no. 12, pp. 1859–1870, Jun. 2022, doi: 10.1177/13524585221097296.
12. F. D. Lublin et al., "Defining the clinical course of multiple sclerosis," *Neurology*, vol. 83, no. 3, pp. 278–286, May 2014, doi: 10.1212/wnl.0000000000000560.
13. J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis," *Neurology*, vol. 33, no. 11, p. 1444, Nov. 1983, doi: 10.1212/wnl.33.11.1444.
14. "Detection of Philippine rice plant diseases: A RESNet50 CNN Approach," *IEEE Conference Publication | IEEE Xplore*, Aug. 17, 2024. <https://ieeexplore.ieee.org/abstract/document/10690946>
15. "Performance of egg sexing classification models in Philippine native duck," *IEEE Conference Publication | IEEE Xplore*, Aug. 07, 2021. <https://ieeexplore.ieee.org/abstract/document/9515279>
16. "Classification of Philippine soybean variety using image processing technique and machine learning method," *IEEE Conference Publication | IEEE Xplore*, Jun. 29, 2024. <https://ieeexplore.ieee.org/abstract/document/10649883>
17. Y. Zhang 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*, vol. 92, no. 9, pp. 861–871, Sep. 2016, doi: 10.1177/0037549716666962.
18. "Multiple sclerosis detection based on biorthogonal wavelet transform, RBF kernel principal component analysis, and logistic regression," *IEEE Journals & Magazine | IEEE Xplore*, 2016. <https://ieeexplore.ieee.org/abstract/document/7747672>
19. Shoeibi A., Khodatars M., Jafari M., Moridian P., Rezaei M., Alizadehsani R., Khozimeh F., Gorriz J.M., Heras J., Panahiazar M., et al. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Comput. Biol. Med.* 2021;136:104697. doi: 10.1016/j.compbiomed.2021.104697



## International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

20. Arani L.A., Hosseini A., Asadi F., Masoud S.A., Nazemi E. Intelligent computer systems for multiple sclerosis diagnosis: A systematic review of reasoning techniques and methods. *Acta Inform. Med.* 2018;26:258–264. doi: 10.5455/aim.2018.26.258-264.
21. Sarbaz Y., Pourakbari H., Vojudi M.H., Ghanbari A. Introducing a decision support system for multiple sclerosis based on postural tremor: A hope for separation of people who might be affected by multiple sclerosis in the future. *Biomed. Eng. Appl. Basis Commun.* 2017;29:1750046. doi: 10.4015/S1016237217500466.
22. Jain, S., Rajpal, N. & Yadav, J. Multiple sclerosis identification based on ensemble machine learning technique. In *Proceedings of the 2nd International Conference on IoT, Social, Mobile, Analytics & Cloud in Computational Vision & Bio-Engineering (ISMAC-CVB 2020)* (2020).





INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | [ijmrset@gmail.com](mailto:ijmrset@gmail.com) |

[www.ijmrset.com](http://www.ijmrset.com)