tvst

Article

Analysis of Multifocal Visual Evoked Potentials Using Artificial Intelligence Algorithms

Samuel Klistorner[1,*], Maryam Eghtedari[1,*], Stuart L. Graham[2], and Alexander Klistorner[1,2]

[1] Save Sight Institute, Sydney Medical School, University of Sydney, Sydney, New South Wales, Australia
[2] Faculty of Medicine and Health Sciences, Macquarie University, Sydney, New South Wales, Australia

Correspondence: Alexander Klistorner, Save Sight Institute, University of Sydney, 8 Macquarie Street, Sydney, NSW 2000, Australia. e-mail: sasha@eye.usyd.edu.au

Received: February 17, 2021
Accepted: December 5, 2021
Published: January 10, 2022

Keywords: visual evoked potential; computational modeling; clinical trial; autoimmune response/disease; anti-inflammatory agents

Citation: Klistorner S, Eghtedari M, Graham SL, Klistorner A. Analysis of multifocal visual evoked potentials using artificial intelligence algorithms. Transl Vis Sci Technol. 2022;11(1):10, https://doi.org/10.1167/tvst.11.1.10

Purpose: Clinical trials for remyelination in multiple sclerosis (MS) require an imaging biomarker. The multifocal visual evoked potential (mfVEP) is an accurate technique for measuring axonal conduction; however, it produces large datasets requiring lengthy analysis by human experts to detect measurable responses versus noisy traces. This study aimed to develop a machine-learning approach for the identification of true responses versus noisy traces and the detection of latency peaks in measurable signals.

Methods: We obtained 2240 mfVEP traces from 10 MS patients using the VS-1 mfVEP machine, and they were classified by a skilled expert twice with an interval of 1 week. Of these, 2025 (90%) were classified consistently and used for the study. ResNet-50 and VGG16 models were trained and tested to produce three outputs: no signal, up-sloped signal, or down-sloped signal. Each model ran 1000 iterations with a stochastic gradient descent optimizer with a learning rate of 0.0001.

Results: ResNet-50 and VGG16 had false-positive rates of 1.7% and 0.6%, respectively, when the testing dataset was analyzed ($n = 612$). The false-negative rates were 8.2% and 6.5%, respectively, against the same dataset. The latency measurements in the validation and testing cohorts in the study were similar.

Conclusions: Our models efficiently analyze mfVEPs with $<2\%$ false positives compared with human false positives of $<8\%$.

Translational Relevance: mfVEP, a safe neurophysiological technique, analyzed using artificial intelligence, can serve as an efficient biomarker in MS clinical trials and signal latency measurement.

Introduction

Multiple sclerosis (MS) is a common neuroinflammatory disorder affecting 2.2 million people globally. It is a devastating diagnosis, often with onset in young adulthood.[1] Remyelination strategies have the potential to revolutionize therapy for this disease. Visual evoked potentials (VEPs), which are scalp electrode recordings of the brain signal in response to a visual stimulus, have been suggested as promising biomarkers for monitoring responses to remyelination therapy in MS clinical trials. To date, the RENEW, RENEWED, SYNERGY, ReBUILD, and stem cell therapy clinical trials have used VEPs for measuring study outcomes, including responses to treatment.[2–5] However, the clinical usefulness of conventional full-field VEPs is limited by the fact that they provide a summed response of all neuronal elements stimulated. As a result, this technique is prone to phase cancellation of dipoles oriented in opposite directions and subsequent loss of potentially clinically relevant information. In addition, it is greatly dominated by the macular region due to its cortical overrepresentation.[6,7]

The multifocal VEP (mfVEP) technique is superior to conventional full-field VEPs in evaluating the integrity of the visual system, because it (1) provides an independent measurement of multiple segments of visual field, thus allowing more accurate detection of smaller defects; (2) eliminates, or at least greatly

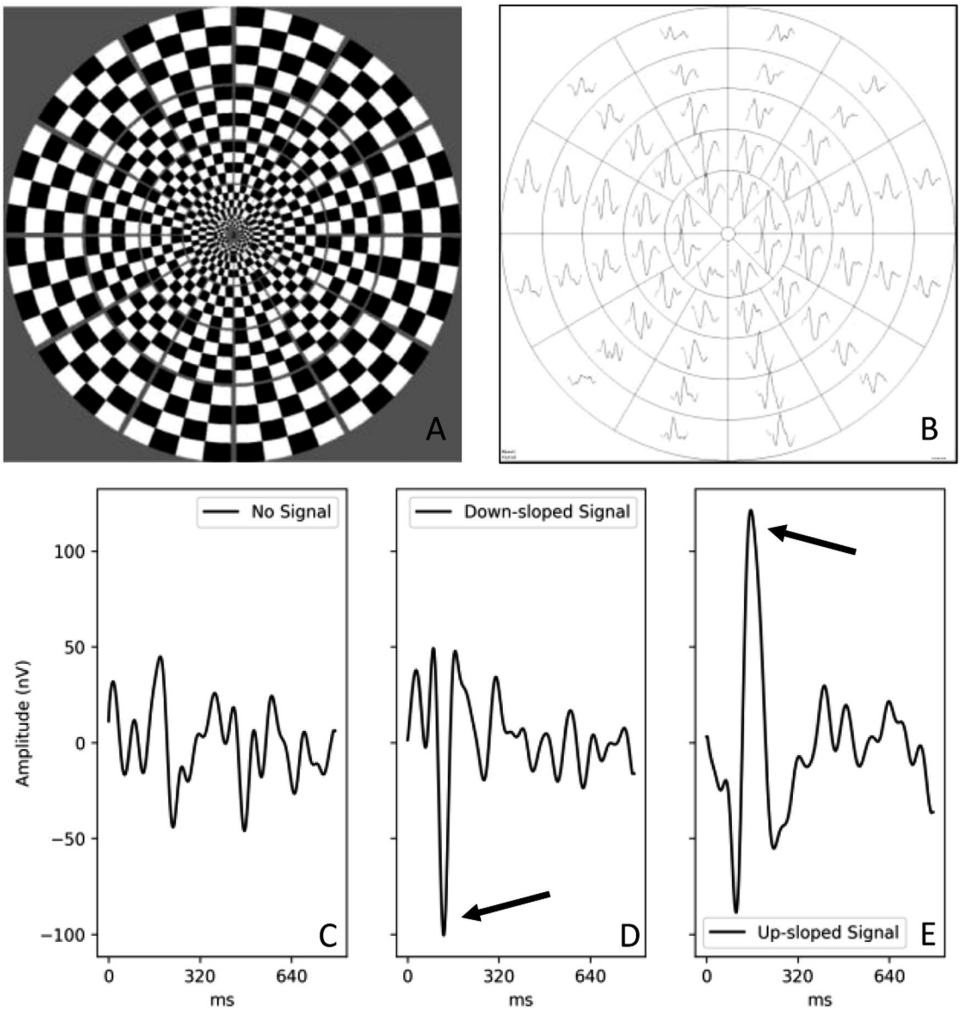

Figure 1. (A) An example of the visual stimulus presented to the subjects. (B) An example of mfVEP responses for one eye (56 traces) across the vertical plane. (C) An example of no signal or a noisy signal. (D) An example of a down-sloped signal. (E) An example of an up-sloped signal.

diminishes, the dipole orientation cancellation effect; and (3) allows assessment of contribution from the more peripheral parts of the visual field.[6,8,9] This advantage is due to simultaneous and independent stimulation of multiple locations of the visual field and recording across two perpendicularly oriented channels using four scalp electrodes.

The mfVEP, however, produces a large amount of data per patient (i.e., each mfVEP recording contains 224 individual traces), which, for accurate analysis of small changes, must be analyzed manually by skilled researchers. The analysis of data typically involves separation of noisy traces (which is a source of high variability) from true (readable) responses and the detection of latency peaks in the true responses. These tasks are onerous and require a high level of training and expertise. As the number of clinical trials and patients in trials grows, leading to larger datasets, the advent of a scalable alternative becomes vital for the speedy generation of results and the incorporation of mfVEPs into both trials and clinical practice. Artificial intelligence (AI) techniques are promising tools to overcome this barrier. Therefore, the aim of the current study was to develop a machine-learning approach for the identification of noisy traces (referred to here as "no-signal") versus measurable mfVEP responses (referred to here as "signal") and the detection of latency peaks in the signals (Fig. 1).

Methods

Standard protocol approvals, registrations, and patient consents were obtained. The study was approved by University of Sydney and Macquarie

University Human Research ethics committees and followed the tenets of the Declaration of Helsinki. Written informed consent was obtained from all participants.

Subjects

Ten consecutive patients with relapsing remitting MS, defined according to the revised McDonald 2010 criteria, and with a history of unilateral optic neuritis of more than 6 months were enrolled.[8] All patients had visual acuity of ≥6/12.

MfVEP Recordings

MfVEPs were recorded monocularly across vertical and horizontal planes using a Visionsearch (Sydney, Australia) system with standard stimulus conditions as described previously.[9] In brief, four gold-disc electrodes (Grass Instruments, West Warwick, RI) were used for bipolar recording with two electrodes positioned 4 cm on either side of the inion, one electrode 2.5 cm above, and another 4.5 cm below the inion in the midline. Electrical signals were recorded along two channels, measured as the difference between superior and inferior electrodes (vertical channel) and between the left and right electrodes (horizontal channel). Fifty-six segments per channel were generated: eight segments in the inner ring and 12 segments for each of the four outer rings (Fig. 1). The size of the segments in individual rings was cortically scaled. In total for this study, 2240 mfVEP traces were analyzed. Data for the training and testing of AI networks was deliberately collected from three different sites, and the analyses were performed by different technicians.

The traces were classified manually twice, a week apart, by a skilled expert (AK) to separate traces with recognizable mfVEP responses ("signal") from noisy traces ("no-signal"), as well as to determine the latency peak of the signal. A signal was further classified into two different configurations: up-sloped or down-sloped. This classification was used to identify the latency of the mfVEP response. The algorithm was based on the identification of positive peaks in cases of an up-sloped configuration of the main response or negative peaks in cases of a down-sloped configuration of the main response (Fig. 1, arrows). In order to ensure good quality of the data, only the segments that were consistently identified by the expert as signal or no-signal in both readings were further used in the study (Fig. 1).

The images of the VEP traces were then used as input to the AI model. No augmentation, transformation, or signal processing was done to the images. The images of traces were divided into two sets: the first set (70% of images) was used for training and validation (training dataset), and the second set (30% of images) was used for testing (testing dataset). The training dataset was further split into 80% for training and 20% for validation. In order to test models against new data, we ensured that traces in the training and testing datasets were from different subjects.

Model Description

Two image-based models were tested: ResNet-50[10] and VGG16.[11] An image-based model was used, as processing the two-dimensional (image) shapes of the mfVEPs closely represents the assessment done by human experts when determining signal versus no signal traces. The models have been proven to be effective in many image classification challenges in the computer science and medical fields. These models were loaded with pre-trained weights leveraging their image recognition detection network.

In this study, each trace was converted to a black-and-white image with a resolution of 540 × 400 pixels. To feed the images into the models, the images were resized to 244 × 244 pixels. The ResNet-50 model was adjusted to take the black-and-white image (rather than a red, green, and blue color image) and output three classes: no signal, up-sloped signal, or down-sloped signal. The VGG16 model input was not adjusted to black-and-white images. Each model ran 1000 iterations with a stochastic gradient descent optimizer, which was selected to have a learning rate of 0.0001. For the latency measurements, in the case of an up-sloped signal, the coordination of the highest point marked the latency; in the case of a down-sloped signal, the coordination of the lowest point marked the latency. These were measured using simple statistical software across the training set ($n = 1413$) and the testing set ($n = 613$); the distribution of these result is presented Figure 2.

In order to maintain high accuracy of longitudinal mfVEP analyses, and latency in particular, it is more important not to include noisy traces (which may introduce a high degree of variability) than to miss some of the true "signal" traces. Therefore, for the model to be successful, a low false-positive (FP) rate (i.e., noisy traces classified as real signal) is required. Conversely, a low false-negative (FN) rate (i.e., traces containing real signal but classified as noise), although still desirable, is far less crucial, as these rates are excluded from analyses and therefore do not affect progression results.

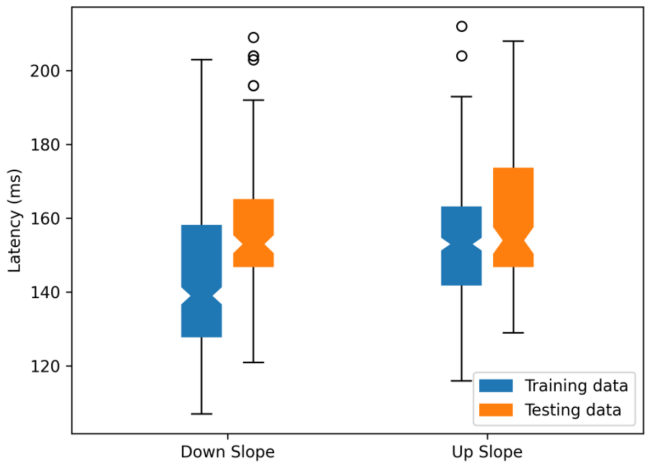

Figure 2. Box plot demonstrating the distribution of latency between down-sloped and up-sloped signals for the training (*blue*) and testing (*orange*) datasets. The lower edge of the box represents the 25th percentile, and the upper edge represents the 75th percentile. The notch is the median. Whiskers are a maximum length of 1.5 × interquartile range. *Circles* represent outliers.

Results

A total of 2240 mfVEP segments were classified twice with an interval of 1 week. Of these, 2025 (or 90%) were classified consistently (i.e., as up-sloped signal, down-sloped signal, or no signal) on both occasions (Table 1). In addition, for cases classified as "signal," the peak was detected in the same location at each reading. These traces were used for training and testing of our AI models. The remaining 215 images were excluded from the study due to discrepancies between the two readings (Table 1). Models were trained using 1413 mfVEP traces (1157 for training, 256 for validation) and tested using the 612 remaining mfVEP traces, of which 249 were originally classified by an expert as "signal" (126 up-sloped signals and 123 down-sloped signals) and 363 as "no signal." A confusion matrix was generated for each of the models for the testing cohort ($n = 612$) for ResNet-50 and VGG16 (Table 2).

Both models demonstrated relatively low rates of FN cases (a situation when down-slope or up-slope signals are classified as "no signal"): 8.2% and 6.5% for ResNet-50 and VGG16, respectively. More importantly, both models displayed very low FP rates ("no signal" classified as either an up-sloped or a down-sloped signal): 1.7% and 0.6% for ResNet-50 and VGG16, respectively.

Both models also performed well in distinguishing between up-sloped and down-sloped Signals. Thus, of the 126 up-sloped signals, both ResNet-50 and VGG16 produced similar results in the validation set, with 103 (81.7%) and 113 (89.7%) correct classifications, respectively. Of the 123 down-sloped signals, 114 and 111 were correctly classified by ResNet-50 and VGG16, resulting in accuracies of 92.7% and 90.2%, respectively. However, by applying a precision metric to ResNet-50 and VGG16, we can see that the detection of up-sloped signals was 97.1% for ResNet-50, and it was 99.1% for VGG16. Similar results were obtained for down-sloped signals, with precision of 97.4% and 99.1%, respectively. Neither of the models misclassified the direction of the slope (e.g., up-sloped classified as down-sloped); all misclassification of signal data could be attributed to signals being classified as no-signal (i.e., FNs). Figure 2 demonstrates similar widespread distributions of latency values for both the training and testing datasets for down-sloped signals and up-sloped signals, indicating the presence of signals with both normal and delayed latency.

Discussion

In this study, we present an approach to the automatic interpretation of mfVEP signals using AI that can provide a rapid and accurate separation of noisy responses from reliable signals, thus enabling change analysis in longitudinal follow-up. This is important, as the mfVEP is one of the few tools available to monitor the state of nerve myelination. Remyelination of chronically demyelinated white matter represents a promising strategy in the treatment of MS. Such an approach offers the potential to prevent accelerated axonal degeneration of damaged (demyelinated) axons from inflammatory mediators and immune effector cells and to restore conduction velocity.[12–14] The validation of remyelinating therapies, however, is hampered by the current lack of consensus

Table 1. Classification of Traces by Human Expert ($N = 2240$ Total Images)

Classification	Percent (%)	*n*
Classified as signal first round but no-signal second round	5.8	131
Classified as no-signal first round but as signal second round	3.8	84
Classified the same between first and second rounds	90.4	2025

Table 2. Classification of Traces by ResNet-50 and VGG16

Classification	Predicted No-Signal	Predicted Up-Sloped Signal	Predicted Down-Sloped Signal	Total
		ResNet-50, *n* (%)		
Actual no-signal	357 (98.3)	3 (0.8)	3 (0.8)	363
Actual up-sloped signal	23 (18.3)	103 (81.7)	0	126
Actual down-sloped signal	9 (7.3)	0	114 (92.7)	123
Total	389 (63.6)	106 (17.3)	117 (19.1)	612
		VGG16, *n* (%)		
Actual no-signal	361 (99.4)	1 (0.02)	1 (0.02)	363
Actual up-sloped signal	13 (10.3)	113 (89.7)	0	126
Actual down-sloped signal	12 (9.8)	0	111 (90.2)	123
Total	386 (63.07)	114 (18.6)	112 (18.3)	612

on use of imaging biomarkers for remyelinating trials, particularly considering the moderate effect of potential remyelinating drugs.[15] Although there are a number of promising therapies, reliable imaging biomarkers for myelin repair remain to be identified. However, evoked potentials and, particularly, mfVEPs due to their ability to directly estimate the speed of axonal conduction are highly sensitive and very accurate quantitative measures of de-/remyelination in both experimental and clinical settings.[2,3,16–18] Although accuracy of VEP measurement and latency, in particular, is essential for monitoring optic nerve function (considering the small degrees of change observed in remyelination trials[3]), the vast amount of mfVEP data typically collected in human clinical trials and manual techniques utilized for latency measurement make it susceptible to error.

AI has been increasingly used in the field of biomedical image analysis. In the current study, we have tested the capability of two image-based AI models to correctly identify the presence of measurable mfVEP traces and separate them from noisy (i.e., unreliable) traces in a group of treated MS patients, among whom we would expect to find both normal and reduced amplitude responses, as well as changes in latency. The patients were therefore representative of the typical clinical scenario where a range of responses may be encountered, even across the field of one recording.

In order to compare AI to human performance, we initially evaluated the ability of an experienced mfVEP analyst to separate identifiable traces with visible responses from noisy (no-signal) traces. The experienced mfVEP reader had an error rate of just below 10%. From an accuracy point of view, it is more important not to overestimate positive responses (not to identify noisy traces as true signals) than to categorize true responses as noise and lose some data. For this reason, when the AI classification was being performed we aimed for a smaller false positive rate (>5%) at the expense of increased FN responses (>10%).

In general, both models demonstrated a high level of precision in identifying measurable traces when the testing dataset was analyzed. False-positive and false-negative rates were well within the expected range (<5% and 10%, respectively). The strong performance of the AI models with regard to achieving results comparable to those of an experienced human analyst is encouraging, particularly considering the time saved when an AI algorithm performs the classification. For example, it takes 20 to 30 minutes to thoroughly analyze mfVEP data for an eye, but the time required for AI to perform the same task is 1 minute. The detection of latency using AI networks to identify the slope of the main signal also demonstrated excellent performance, with no misclassification in the testing dataset.

To the best of our knowledge, this is the first study to propose such an approach for analyzing mfVEPs and its use in clinical practice for MS. Qiao[19] described a deep learning technique based on the VGG19 model that demonstrated an accuracy of 90.6% when analyzing data from patients with suprasellar tumors. In our study, both models demonstrated high precision of over 97% for signal detection and a FP rate of less than 2%. Given that the traces were recorded using different operators at different sites, the result of classification is not site (or operator) specific; however, the same model of mfVEP machine (OV-1) was used to obtain all of the recordings. Hence, it remains to be seen how well the algorithm will perform when applied to data collected using different models of mfVEP machines.

The primary aim of the current study was to identify and remove noisy (unmeasurable) mfVEP traces from the analysis of the latency progression. Because MS patients are known to exhibit the entire range of mfVEP waveforms, data from normal controls were not required in this modeling. However, the current approach can now be applied to a population of normal subjects to determine the overall specificity of mfVEP.

Both models proved to be efficient algorithms in detecting mfVEPs and were able to efficiently replicate or outperform a human. The processing times for both models were similar. However, as demonstrated in Table 2, VGG16 has higher acuity in detecting true signals and maintained a low FP rate. As discussed earlier, this is an important point when considering correct latency measurements and subsequent applications in clinical decision making; thus, the authors recommend that this model be used in clinical trials versus ResNet-50. As a future direction for this study, an AI model processing one-dimensional numerical data could be considered to complement the findings from our study, which used image-based, two-dimensional AI modeling of mfVEPs.

In conclusion, the application of AI to mfVEP analysis to separate true traces from those contaminated by noise and to identify latency peaks proved to be accurate, reliable, and efficient. It opens up new possibilities for using mfVEPs as biomarkers in clinical trials of remyelinating agents, as these will monitor latency changes over time.[9] This tool can be used in clinical practice and provide a fast and relatively low-cost assessment of the remyelinating capacity of new therapies in MS.[20]

Acknowledgments

Supported by grants from the National Multiple Sclerosis Society (RG4716A6/3), the Sydney Eye Hospital Foundation, the Claffy Foundation, and Sydney Medical School Foundation (K6602/RY285).

Disclosure: **S. Klistorner**, None; **M. Eghtedari**, None; **S.L. Graham**, None; **A. Klistorner**, None

* SK and ME contributed equally to this work.

References

1. Wallin MT, Culpepper WJ, Nichols E, et al. Global, regional, and national burden of multiple sclerosis 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2019;18:269–285.
2. Cadavid D, Balcer L, Galetta S, et al. Safety and efficacy of opicinumab in acute optic neuritis (RENEW): A randomised, placebo-controlled, phase 2 trial. *Lancet Neurol*. 2017;16:189–199.
3. Green AJ, Gelfand JM, Cree BA, et al. Clemastine fumarate as a remyelinating therapy for multiple sclerosis (ReBUILD): A randomised, controlled, double-blind, crossover trial. *Lancet*. 2017;390:2481–2489.
4. Plemel JR, Liu W, Yong VW. Remyelination therapies: A new direction and challenge in multiple sclerosis. *Nat Rev Drug Discov*. 2017;16:1–18.
5. Connick P, Kolappan M, Crawley C, et al. Autologous mesenchymal stem cells for the treatment of secondary progressive multiple sclerosis: An open-label phase 2a proof-of-concept study. *Lancet Neurol*. 2012;11:150–156.
6. Alshowaeir D, Yiannikas C, Klistorner A. Multifocal visual evoked potential (mfVEP) and pattern-reversal visual evoked potential changes in patients with visual pathway disorders: a case series. *Neuroophthalmology*. 2015;39:220–233.
7. Klistorner A, Fraser C, Garrick R, Graham SL, Arvind H. Correlation between full-field and multifocal VEPs in optic neuritis. *Doc Ophthalmol*. 2008;116:19–27.
8. Klistorner AI, Graham SL, Grigg JR, Billson FA. Multifocal topographic visual evoked potential: improving objective detection of local visual field defects. *Invest Ophthalmol Vis Sci*. 1998;39:937–950.
9. Klistorner A, Chai Y, Leocani L, Albrecht P, Aktas O, Butzkueven H. Assessment of opicinumab in acute optic neuritis using multifocal visual evoked potential. *CNS Drugs*. 2018;32:1159–1171.
10. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *arXiv*. 2015, arXiv:1512.03385v1.
11. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. 2015, arXiv:1409.1556v6.
12. Bruck W, Kuhlmann T, Stadelmann C. Remyelination in multiple sclerosis. *J Neurol Sci*. 2003;206:181–185.
13. Kornek B, Storch MK, Weissert R, et al. Multiple sclerosis and chronic autoimmune encephalomyelitis: a comparative study of axonal injury in active, inactive and remyelinated lesions. *Am J Pathol*. 2000;157:267–276.

14. De Groot CJA, Ruuls SR, Theeuwes JMW, Dijkstra CD, Van der Valk P. Immunocytochemical characterization of the expression of inducible and constitutive isoforms of nitric oxide synthase in demyelinating multiple scleroisis lesions. *J Neuropathol Exp Neurol*. 1997;56:10–20.
15. Oh J, Ontaneda D, Azevedo C, et al. Imaging outcome measures of neuroprotection and repair in MS: a consensus statement from NAIMS. *Neurology*. 2019;12:519–533.
16. Heidari M, Radcliff AB, McLellan GS, et al. Evoked potentials as a biomarker of remyelination. *Proc Natl Acad Sci USA*. 2019;116:27074–27083.
17. Van Der Walt A, Kolbe S, Mitchell P, et al. Parallel changes in structural and functional measures of optic nerve myelination after optic neuritis. *PLoS One*. 2015;10:e0121084.
18. You Y, Klistorner A, Thie J, Graham S. Latency delay of visual evoked potential is a real measurement of demyelination in a rat model of optic neuritis. *Invest Ophthalmol Vis Sci*. 2011;52:6911–6918.
19. Qiao N. Using deep learning for the classification of images generated by multifocal visual evoked potential. *Front Neurol*. 2018;9:638.
20. Klistorner A, Triplett JD, Barnett MH, et al. Latency of multifocal visual evoked potential in multiple sclerosis: A visual pathway biomarker for clinical trials of remyelinating therapies. *J Clin Neurophysiol*. 2021;38:186–191.