Harmonic Kernel Mixing for Interpretable Oculomotor Event Segmentation in Video-nystagmography

Kunal Chaturvedi[1], Ali Braytee[1], Nicholas Yang[2], Imelda Hannigan[2], Nicole Reid[3], Andrew Bradshaw[3], Miriam Welgampola[2,3], Mukesh Prasad[1]

[1] School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia
[2] Central Clinical School, University of Sydney, Sydney, Australia
[3] Institute of Clinical Neurosciences, Royal Prince Alfred Hospital, Sydney, Australia
kunal.chaturvedi@uts.edu.au; mukesh.prasad@uts.edu.au

Abstract. Nystagmus is an involuntary rhythmic eye movement that is central to the diagnosis of vertigo and related neurological disorders. It consists of slow phases, reflecting pathological ocular drift, and fast phases, corrective saccades that reset gaze. Quantification of slow-phase velocity (SPV) is clinically essential, providing an objective and interpretable measure for physiological verification of nystagmus dynamics in vestibular assessment. However, traditional SPV estimation typically relies on threshold-based detection and hand-tuned heuristics that can be sensitive to noise, artifacts, and inter-subject variability. We present an end-to-end framework for oculomotor event segmentation that directly supports SPV reconstruction from predicted slow phases. Our key contribution is Harmonic Kernel Mixing, an interpretable spectral front-end that augments a Sinc-parameterized band-pass filterbank with a learnable harmonic cross-band coupling among related frequency bands while retaining transparent cut-off frequencies and coupling weights. Building on this front-end, we introduce a spectro-temporal network with multi-scale temporal encoding to segment slow phase, fast phase, no nystagmus, and artifact regions. On the Omniax clinical dataset with patient-wise stratified 5-fold cross-validation, our method achieves 0.735±0.005 macro-F1, the highest mean macro-F1 among evaluated time-series baselines. We further evaluate on GazeCom and an independent set of 164 BPV cases, where predicted slow-phase segments enable accurate SPV reconstruction (MAE 2.44±2.99 °/s) with <2 s mean absolute error for key SPV temporal landmarks.

Keywords: Video-nystagmography, Vertigo, Nystagmus, Temporal Analysis.

1 Introduction

Vertigo is a false perception of motion arising from dysfunction of the vestibular system and is a frequent cause of presentation to acute and primary care. When diagnosis and treatment are delayed, vertigo is associated with substantial morbidity and societal cost, including falls and injury risk, and increased healthcare utilization [1]. Clinical assessment relies on nystagmus, a key biomarker of vestibular imbalance. While experts may observe nystagmus with the naked eye, video-nystagmography (VNG) provides the gold standard, using vision-denied infrared recording to capture the precise

characteristics essential for differentiating peripheral mechanical disorders from central pathologies [2].

For diagnosis, however, the mere presence of nystagmus is insufficient; critical to accurate interpretation is the quantitative analysis of the slow phase, which reflects the pathological drift of the eye caused by vestibular imbalance. The velocity of this slow phase (SPV) tracks the intensity of the underlying neural or mechanical stimulation over time. Consequently, the morphological characteristics of the SPV profile serve as a vital diagnostic clue. For instance, in patients with Benign Positional Vertigo (BPV), a crescendo-decrescendo velocity profile serves as a "mechanical fingerprint" for canalolithiasis, distinguishing it from the constant velocity decay typical of cupulolithiasis or central pathologies that may mimic peripheral nystagmus direction [3].

Despite its diagnostic importance, extracting reliable SPV profiles from VNG recordings remains a significant clinical bottleneck. Current interpretation relies on manual inspection of the nystagmus traces to remove artifacts (such as blinks) [4–6], accurately segment slow and fast phases, and calculate velocity slopes. Performing this process manually is labor-intensive and suffers from high inter-rater variability. Recent learning-based methods [7–10] have sought to automate VNG analysis, but most approaches focus on high-level classification rather than the granular phase segmentation needed to derive SPV continuously. This creates a mismatch between model outputs and clinical verification: a label alone offers limited support for mechanism-level reasoning when clinicians expect an interpretable velocity profile. Conversely, classical signal-processing pipelines [11, 12] can support slow-/fast-phase analysis, but typically rely on hand-crafted preprocessing and threshold-based heuristics, and may be sensitive to noise, baseline drift, and recording variability common in bedside settings.

This paper addresses automated VNG analysis from the perspective of clinically interpretable quantification. Specifically, we aim to produce granular segmentation of VNG-derived eye-position traces, discriminating between slow phases, fast phases, artifacts, and periods of no nystagmus, such that SPV can be directly derived and inspected downstream. A central challenge in achieving this robust separation is the spectral complexity of oculomotor signals [13, 14]. These signals are non-stationary and often exhibit energy distributed across multiple related frequency components, with variability introduced by noise, baseline drift, and intermittent artifacts [15, 16]. Standard convolutional networks learn unconstrained filters that are difficult to interpret in the frequency domain and may be sensitive to such variability. SincNet [17] improves interpretability by restricting filters to band-pass shapes, but it predominantly treats each frequency band in isolation, which can limit the model's ability to exploit structured relationships across frequencies that recur in physiological eye-movement patterns.

To capture the complex, non-stationary nature of oculomotor signals while preserving interpretability in time–frequency space, we introduce a Harmonic Kernel Mixing (HKM) spectral front-end. HKM injects a learnable harmonic prior that softly couples neighboring band-pass filters, extracting coherent spectral structures rather than treating frequency bands independently. Combined with multi-scale temporal modeling, we present NyST-Net, a spectro-temporal architecture to deliver oculomotor event

segmentation, enabling automated derivation of continuous SPV profiles for physiological verification. Our primary contributions are as follows:

- This paper proposes NyST-Net, a hybrid architecture that integrates HKM with a multi-scale temporal backbone to perform oculomotor event segmentation, enabling automated derivation of continuous SPV profiles from predicted slow-phase segments.
- HKM introduces a novel, interpretable spectral front-end that injects a harmonic prior to softly couple Sinc-parameterized band-pass filters with harmonically related neighboring bands.
- We evaluate NyST-Net on the Omniax clinical dataset against representative baselines, and further benchmark its performance on the public GazeCom dataset.

2 Methods

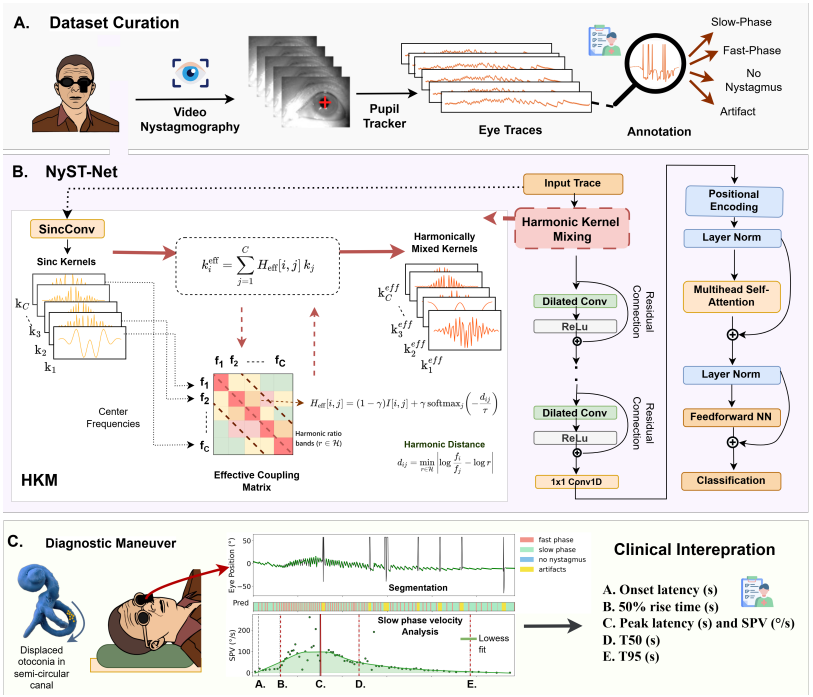

Fig. 1. Overview of the proposed pipeline. A. From VNG recordings, eye traces are extracted and event labels are annotated. B. Architecture of NyST-Net, comprising HKM, and a multiscale temporal backbone to segment regions. C. Predicted slow-phase segments are post-processed to reconstruct SPV profiles and compute landmarks.

2.1 NyST-Net for Oculomotor Event Segmentation

Given an extracted eye-position trace $x \in \mathrm{R}^T$, we predict a per-frame label sequence for oculomotor events (slow phase, fast phase, no nystagmus, artifact). This is achieved

through NyST-Net (Fig. 1B), which consists of an interpretable spectral front-end based on Harmonic Kernel Mixing (HKM), followed by hierarchical temporal modelling. To extract robust spectral features from non-stationary eye position traces while preserving interpretability, HKM augments a learnable Sinc band-pass filterbank with soft, structured cross-band coupling.

We instantiate the spectral front-end as a bank of C Sinc band-pass filters of length K. For channel i, the base kernel is

$$g_i[n] = 2f_{h_i}\mathrm{sinc}(2\pi f_{h_i} n) - 2f_{l_i}\mathrm{sinc}\big(2\pi f_{l_i} n\big), \tag{1}$$

where f_{l_i} and f_{h_i} are the learnable low and high cutoff frequencies, $n \in \left[-\frac{K-1}{2}, \frac{K-1}{2}\right]$ is the discrete time index. A Hamming window is then applied to reduce spectral leakage, yielding base kernels $\{k_i\}_{i=1}^{C}$.

To enable information flow between harmonically related bands, we compute an affinity matrix based on the log-frequency ratio between filter centers. Let the center frequency be $f_i = f_{l_i} + \frac{1}{2}(f_{h_i} - f_{l_i})$. For channels i, j, we define a harmonic distance to a small ratio set $\mathcal{H}$:

$$d_{ij} = \min_{r \in \mathcal{H}} \left| log\left(\frac{f_i}{f_j}\right) - logr \right|, \tag{2}$$

masking ratios that would place rf_j outside the valid band $(f_{min}, f_s/2)$ where f_s is the sampling rate and $f_{min} > 0$. Distances are converted to row-normalized affinities with a learnable temperature $\tau > 0$, and combined with a learnable residual gate $\gamma \in (0,1)$:

$$H_{eff} = (1-\gamma)I + \gamma softmax\left(-\frac{d}{\tau}\right), \tag{3}$$

where softmax is applied row-wise. H_{eff} yields a transparent coupling map where $H_{eff}[i,j]$ quantifies how strongly the i-th effective band borrows from the j-th base band. We then form effective kernels by mixing base kernels and use them to convolve the input trace,

$$k_i^{eff}[n] = \sum_j H_{eff}[i,j]\, k_j[n], \quad y_i = x * k_i^{eff}, \tag{4}$$

Stacking $\{y_i\}$ yields $Y = [y_1, \ldots, y_C] \in \mathrm{R}^{T \times C}$. HKM remains interpretable through (i) the learned cutoff frequencies defining each band-pass kernel and (ii) the explicit coupling weights H_{eff}.

Given the HKM features $Y \in \mathrm{R}^{T \times C}$, we apply a multi-scale temporal backbone to produce contextualized per-frame embeddings. First, a temporal convolutional network (TCN) consisting of residual dilated 1D convolutions preserves frame alignment while increasing the effective receptive field. A final 1×1 projection maps features to width d, producing $H \in \mathrm{R}^{T \times d}$. We then add learnable positional encodings $P \in \mathrm{R}^{T \times d}$ and form $S = H + P$. The sequence S is passed through a Transformer encoder with multi-head self-attention and a position-wise feed-forward block. Finally, a lightweight per-

frame prediction head outputs class logits, yielding frame-level segmentation into the four event classes.

2.2 Slow-phase Profile Reconstruction

To derive an SPV profile from the predicted segmentation, we compute velocity from the eye-position trace by numerical differentiation with respect to time. For each predicted slow-phase segment, SPV is estimated as the median absolute velocity, and segments shorter than 0.05 s are discarded. Segment-level SPV values are then smoothed using LOWESS to obtain a continuous SPV envelope. We evaluate absolute SPV envelopes for landmark recovery; signed SPV can be computed from the same predicted slow-phase intervals for direction-aware clinical review. From this, we extract landmarks used to summarize the maneuver-evoked response driven by otoconia-induced endolymph flow (Fig. 1C). Onset latency is defined as the time from maneuver onset to the first sustained rise of the envelope above a small fraction of its peak. Peak SPV is the maximum of the envelope, and peak latency is the time from maneuver onset to this maximum. The 50% rise time is the interval from onset to the first crossing of 50% of peak. The decay times T50 and T95 are defined as the times after peak when the envelope first falls below 50% and 5% of peak, respectively.

3 Experiments and Results

3.1 Implementation Details

Datasets. We evaluate NyST-Net on two eye-movement datasets with distinct acquisition conditions. Omniax is a private clinical dataset comprising 3,153 eye movement traces from 665 patients, collected retrospectively using the Epley Omniax Rotator at an outpatient facility. Clinical data use was approved by the Sydney Local Health District Human Research Ethics Committee (X18-0087; 2019/ETH06926). Monocular eye videos were recorded at 30 Hz, and 2D eye-position traces were extracted. Recordings include Hallpike and roll maneuver variants to elicit provoked vestibular nystagmus. Frame-level ground truth was annotated into four classes: slow phase, fast phase, no nystagmus, and artifact (e.g., blinks/tracking loss). Recordings were clinically reviewed before event labels were created by a trained engineer; the labelled frames comprised slow phase 54.3%, fast phase 16.2%, no nystagmus 21.2%, and artifact 8.2%. In addition, we benchmark on the GazeCom dataset [18], which contains eye-movement recordings from 54 participants acquired at 250 Hz during naturalistic viewing, with labels for fixation, saccade, smooth pursuit, and noise.

Experimental Settings. We employ patient-wise 5-fold cross-validation (CV), using four folds for development and one fold for testing; within each development split, we reserve 10% of training patients for validation. Dataset-specific settings and hyperparameters were selected using development/validation data only; test folds were not used

for hyperparameter selection, model selection, or early stopping. Models are implemented in PyTorch and optimized via AdamW (LR 10^{-3}, weight decay 10^{-4}) for 200 epochs, with a batch size of 2048, using cross-entropy loss, early stopping (patience of 20) and a reduce-on-plateau scheduler. For Omniax, we treat the horizontal and vertical components as single-channel inputs, keeping both components from each recording within the same patient-wise fold.

During training, inputs were segmented into 300-sample windows with stride 5; validation and test evaluation used non-overlapping windows for macro- and class-wise F1. NyST-Net uses 32 SincConv filters ($K = 51$) with HKM harmonic ratio set tuned to $\mathcal{H} = \{2/3,1,3/2\}$, where τ and γ are learned end-to-end, followed by a two-block TCN (64 and 96 channels; dilations 1 and 2) and a 1-layer Transformer encoder (2 heads, dropout 0.15). For GazeCom, we maintain identical data preprocessing and evaluate using the contiguous temporal CV protocol of [19]. Inputs are segmented into 250-sample windows (stride 10), and NyST-Net uses SincConv (K=31) with similar harmonic ratio set, a deeper TCN stack (128 and 256 channels; dilations 1, 2, and 4), and a 4-layer Transformer (8 heads, dropout 0.25). For Omniax baselines, we ensure a fair comparison by using the same patient splits, windowing procedure, and training protocol, and selecting models based on the validation patients within each outer fold.

3.2 Comparison with Baseline Methods

Table 1 compares NyST-Net with two baseline families on Omniax: a classical VOG phase-detection pipeline [11], and representative deep time-series backbones [20–22]. Since [10] is formulated for slow-/fast-phase detection and does not define artifact or no-nystagmus categories, we report only these two classes. NyST-Net achieves the highest mean F1 for both slow and fast phases, indicating improved phase-event segmentation on noisy VNG traces. The gain is especially notable for fast phases, where rule-based methods are more susceptible to noise, baseline drift, and brief high-frequency transitions. Across the learned baselines, NyST-Net also attains the highest mean F1 for the no-nystagmus class, while artifact performance remains comparable to strong deep backbones. Overall, these results support NyST-Net as a strong and clinically relevant choice for phase segmentation in noisy VNG traces.

Table 1. Performance comparison of NyST-Net and baseline models on the Omniax dataset using F1 (mean±std)

Model	Slow phase	Fast phase	No nystagmus	Artifact
Winnick et al. [11]	0.699±0.017	0.132±0.003	-	-
ResNet1D [20]	0.817±0.016	0.755±0.010	0.432±0.027	0.775±0.024
InceptionTime [21]	0.830±0.017	0.766±0.007	0.477±0.028	**0.777±0.019**
Conformer [22]	0.832±0.014	0.735±0.012	0.555±0.045	0.762±0.026
NyST-Net (HKM)	**0.840±0.014**	**0.769±0.004**	**0.563±0.019**	0.765±0.018

3.3 Clinical Verification on Benign Positional Vertigo

To assess clinical utility, we evaluated NyST-Net's ability to reconstruct SPV profiles on an independent Omniax cohort of 164 BPV recordings. We computed the reference SPV using the same procedure applied to the ground-truth slow-phases. As shown in Fig. 2, the reconstructed SPV closely matched the reference envelope, achieving an MAE of 2.44±2.99 °/s and an RMSE of 4.14±5.36 °/s. The model also recovered key diagnostic landmarks with small timing error (onset latency 0.51±0.85 s; peak latency 1.16±1.49 s). While T95 decay exhibited higher variability due to the difficulty of segmenting low-amplitude tails, the primary SPV envelope trends were retained, supporting NyST-Net as a front-end for automated SPV quantification in clinical VNG. The full inference-to-SPV pipeline required 0.011 ± 0.005 s per recording on an NVIDIA L40 GPU.

Fig. 3 illustrates a representative BPV test patient, showing both phase segmentation and the resulting SPV. The predicted SPV trajectory tracks the reference trajectory with MAE of 1.04°/s. This behavior is consistent with the HKM coupling map (Fig. 3, Right). In the HKM panel, high coupling mass concentrates near the harmonic reference lines, indicating that the representation preferentially links filter pairs whose frequency ratios match the harmonic set. In this example, coupling is most pronounced in the low-frequency range, consistent with dominant periodic components in the nystagmus trace. Isolated transient spikes show weaker alignment with the harmonic coupling pattern and are labeled as artifacts, which may help maintain stable slow-phase segmentation without distorting the surrounding SPV envelope.

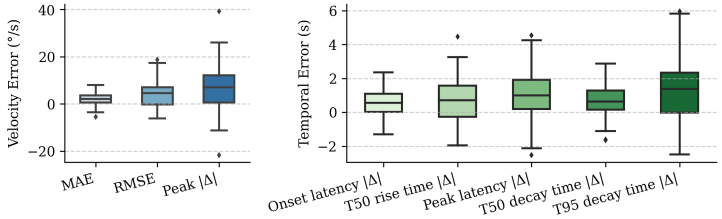

Fig. 2. Box plots show aggregate error distributions for (Left) velocity metrics and (Right) temporal metrics for SPV reconstruction across 164 BPV profiles.

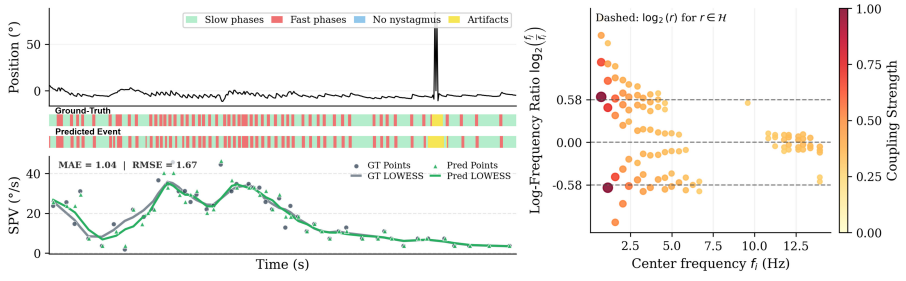

Fig. 3. Qualitative visualization of NyST-Net outputs on a representative BPV trace. (Left) Raw eye-position trace (top), predicted events compared to ground truth (middle), and the reconstructed SPV trajectory (bottom). (Right) HKM coupling map with harmonic reference lines.

3.4 External Evaluation on GazeCom

Table 2 summarizes GazeCom baselines with aligned label definitions and reported sample- and event-level F1 metrics. As shown in Table 2, NyST-Net demonstrates competitive performance on GazeCom, particularly on event-level metrics. While some specialized architectures like Skip-AttSeqNet [23] report stronger sample-level scores for fixation and saccade, NyST-Net attains strong event-level performance across classes, including smooth pursuit (SP) 0.725 mean F1 and saccades with 0.980 mean F1. These results indicate that the proposed spectro-temporal design transfers to a distinct eye-movement domain with different acquisition settings, providing an external benchmark beyond the clinical Omniax cohort.

Table 2. Performance comparison on the GazeCom dataset using F1 (mean)

Model	Sample				Event			
	Fixation	Saccade	SP	Noise	Fixation	Saccade	SP	Noise
1DCNN–BLSTM [18]	0.939	0.893	0.703	—	0.898	0.947	0.596	—
TCNs [23]	0.942	0.895	0.73	—	0.819	0.905	0.421	—
I-BDT [19]	0.868	0.510	0.422	—	0.682	0.472	0.377	—
CNN-BiLSTM [19]	0.884	0.821	0.406	0.572	0.908	0.958	0.485	0.760
CNN-LSTM [19]	0.881	0.822	0.390	0.579	0.903	0.956	0.459	0.641
TCN [19]	0.920	0.839	0.573	0.596	0.936	0.972	0.611	0.774
Skip-AttSeqNet [23]	**0.943**	**0.899**	**0.739**	—	0.917	0.95	0.691	—
NyST-Net (Ours)	0.935	0.867	0.715	0.688	**0.953**	**0.980**	**0.725**	0.791

3.5 Ablation Study

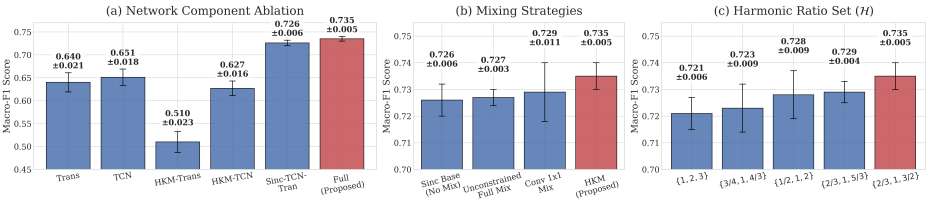

Fig. 4. Ablation of the proposed framework using macro F1. (a) Network component analysis. (b) Comparison of front-end mixing strategies. (c) Evaluation of various harmonic ratio sets.

To determine the optimal configuration, we ablated NyST-Net components and HKM design choices on Omniax. Fig. 4a shows that temporal modeling alone underperforms configurations that include the spectral front-end, indicating that explicit spectral modeling is important for robust frame-wise oculomotor event segmentation.

To verify that gains are not attributable to generic cross-channel mixing, we compared HKM against matched controls applied to the same Sinc filterbank: (i) a learned unconstrained full mixing matrix, (ii) pointwise 1×1 channel mixing, and (iii) Sinc-only without mixing. HKM achieved the best performance among these variants, supporting the benefit of the proposed harmonic prior beyond generic channel mixing (Fig. 4b). We also performed a sensitivity analysis over alternative ratio sets (Fig. 4c) and found

that $\mathcal{H} = \{2/3,1,3/2\}$ consistently yielded the strongest performance among the alternatives (0.721–0.729).

4 Conclusion

We introduced NyST-Net, built around Harmonic Kernel Mixing, an interpretable spectral coupling mechanism, for oculomotor event segmentation. By injecting a structured harmonic prior while retaining explicit cutoff frequencies and coupling weights, HKM promotes robust feature extraction from non-stationary VNG signals. Evaluated on the Omniax clinical cohort and benchmarked on the public GazeCom, NyST-Net achieved the highest mean segmentation performance among representative baselines. On an independent BPV cohort, the resulting slow-phase predictions enabled accurate reconstruction of SPV envelopes and recovery of key diagnostic temporal landmarks. Future work will prioritize prospective multi-center validation and broader clinical deployment studies.

Disclosure of Interests. The authors declare no competing interests.

References

1. Prell, T., Finn, S., Axer, H.: How Healthcare Utilization Due to Dizziness and Vertigo Differs Between Older and Younger Adults. Front. Med. 9, 852187 (2022). https://doi.org/10.3389/fmed.2022.852187.
2. Friedrich, M.U., Schneider, E., Buerklein, M., Taeger, J., Hartig, J., Volkmann, J., Peach, R., Zeller, D.: Smartphone video nystagmography using convolutional neural networks: ConVNG. J Neurol. 270, 2518–2530 (2023). https://doi.org/10.1007/s00415-022-11493-1.
3. Halmágyi, G.M., Akdal, G., Welgampola, M.S., Wang, C.: Neurological update: neuro-otology 2023. J Neurol. 270, 6170–6192 (2023). https://doi.org/10.1007/s00415-023-11922-9.
4. Miles, R., Zapala, D.: Vestibular Function Measurement Devices. Semin Hear. 36, 49–74 (2015). https://doi.org/10.1055/s-0034-1396926.
5. Punuganti, S.A.: Automatic detection of nystagmus in bedside VOG recordings from patients with vertigo, (2019).
6. Yang, N., Chaturvedi, K., Reid, N., Dyball, A.C., Argaet, E.C., Bradshaw, A.P., Wang, C., Rafi, A., Rosengren, S.M., Halmagyi, G.M., Black, D.A., Bharathy, G., Braytee, A., Prasad, M., Welgampola, M.S.: The need for speed: using nystagmus velocity profiles and machine learning models to separate canalithiasis BPV from its mimics. J Neurol. 272, 761 (2025). https://doi.org/10.1007/s00415-025-13403-7.
7. Newman, J.L., Phillips, J.S., Cox, S.J.: 1D Convolutional Neural Networks for Detecting Nystagmus. IEEE J. Biomed. Health Inform. 25, 1814–1823 (2021). https://doi.org/10.1109/JBHI.2020.3025381.
8. Qiu, X., Shi, S., Tan, X., Qu, C., Fang, Z., Wang, H., Gao, Y., Wu, P., Li, H.: Gram-based Attentive Neural Ordinary Differential Equations Network for Video Nystagmography Classification. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV).

pp. 21282–21291. IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.01951.
9. Mun, S.B., Kim, Y.J., Lee, J.H., Han, G.C., Cho, S.H., Jin, S., Kim, K.G.: Deep Learning-Based Nystagmus Detection for BPPV Diagnosis. Sensors. 24, 3417 (2024). https://doi.org/10.3390/s24113417.
10. Nguyen, T.-A.-Q., Hashmi, E., Yamin, M.M., Beghdadi, A., Cheikh, F.A., Ullah, M.: Benign Paroxysmal Positional Vertigo Disorders Classification Using Eye Tracking Data. In: Maglogiannis, I., Iliadis, L., Macintyre, J., Avlonitis, M., and Papaleonidas, A. (eds.) Artificial Intelligence Applications and Innovations. pp. 174–185. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-63215-0_13.
11. Winnick, A.A., Chen, C.-C., Chang, T.-P., Kuo, Y.-H., Wang, C.-F., Huang, C.-H., Yang, C.-C.: Automated nystagmus detection: Accuracy of slow-phase and quick-phase algorithms to determine the presence of nystagmus. Journal of the Neurological Sciences. 442, 120392 (2022). https://doi.org/10.1016/j.jns.2022.120392.
12. Punuganti, S.A., PhD, J.O.-M.: Detection of Saccades and Quick-Phases in Eye Movement Recordings with Nystagmus. In: ACM Symposium on Eye Tracking Research and Applications. pp. 1–5. ACM, Stuttgart Germany (2020). https://doi.org/10.1145/3379156.3391353.
13. Harezlak, K., Kasprowski, P.: Understanding Eye Movement Signal Characteristics Based on Their Dynamical and Fractal Features. Sensors (Basel). 19, 626 (2019). https://doi.org/10.3390/s19030626.
14. Mahanama, B., Jayawardana, Y., Rengarajan, S., Jayawardena, G., Chukoskie, L., Snider, J., Jayarathna, S.: Eye Movement and Pupil Measures: A Review. Front. Comput. Sci. 3, 733531 (2022). https://doi.org/10.3389/fcomp.2021.733531.
15. Mantokoudis, G., Otero-Millan, J., Gold, D.R.: Current concepts in acute vestibular syndrome and video-oculography. Current Opinion in Neurology. 35, 75–83 (2022). https://doi.org/10.1097/WCO.0000000000001017.
16. Abadi, R.V., Akman, O.E., Arblaster, G.E., Clement, R.A.: Analysing nystagmus waveforms: a computational framework. Sci Rep. 11, 9761 (2021). https://doi.org/10.1038/s41598-021-89094-7.
17. Ravanelli, M., Bengio, Y.: Speaker Recognition from Raw Waveform with SincNet. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 1021–1028. IEEE, Athens, Greece (2018). https://doi.org/10.1109/SLT.2018.8639585.
18. Startsev, M., Agtzidis, I., Dorr, M.: 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. Behav Res. 51, 556–572 (2019). https://doi.org/10.3758/s13428-018-1144-2.
19. Elmadjian, C., Gonzales, C., Costa, R.L.D., Morimoto, C.H.: Online eye-movement classification with temporal convolutional networks. Behav Res. 55, 3602–3620 (2022). https://doi.org/10.3758/s13428-022-01978-2.
20. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 1578–1585. IEEE, Anchorage, AK, USA (2017). https://doi.org/10.1109/IJCNN.2017.7966039.
21. Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.-A., Petitjean, F.: InceptionTime: Finding AlexNet for time

series classification. Data Min Knowl Disc. 34, 1936–1962 (2020). https://doi.org/10.1007/s10618-020-00710-y.

22. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition. In: Interspeech 2020. pp. 5036–5040. ISCA (2020). https://doi.org/10.21437/Interspeech.2020-3015.
23. Wang, X., Fan, L., Li, H., Bi, X., Jiang, W., Ma, X.: Skip-AttSeqNet: Leveraging skip connection and attention-driven Seq2seq model to enhance eye movement event detection in Parkinson's disease. Biomedical Signal Processing and Control. 99, 106862 (2025). https://doi.org/10.1016/j.bspc.2024.106862.