Preprints.org

Article

Not peer-reviewed version

Computer Vision for ABR Interpretation: Simultaneous Hearing Threshold and Wave V Latency Detection from Clinical Report Images

Andrzej Sroczyński *, Tomasz Śmiałkowski, Andrzej Czyżewski, Krzysztof Kochanek, Henryk Skarżyński

Posted Date: 30 October 2026

Keywords: auditory brainstem response; hearing threshold detection; wave V latency; object detection; deep learning; computer vision

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Computer Vision for ABR Interpretation: Simultaneous Hearing Threshold and Wave V Latency Detection from Clinical Report Images

Andrzej Sroczyński [1,*], Tomasz Śmiałkowski [1], Andrzej Czyżewski [1], Krzysztof Kochanek [2] and Henryk Skarżyński [2]

[1] Gdansk University of Technology; Faculty of Electronics, Telecommunication and Informatics; Multimedia Systems Department; ul. Narutowicza 11/12, 80-233 Gdansk, Poland

[2] Institute of Physiology and Pathology of Hearing, ul. M. Mochnackiego 10, 02-042 Warszawa, Poland

* Correspondence: andrzej.sroczynski@pg.edu.pl or a.sr@sound.eti.pg.gda.pl

Abstract

Auditory brainstem response (ABR) testing is widely used in hearing screening and clinical diagnostics. In routine practice, ABR recordings are often interpreted from graphical reports rather than raw digital signals, making wave V identification and hearing-threshold estimation challenging, particularly near threshold. This study presents a computer-vision framework that estimates hearing thresholds and wave V latencies directly from ABR report images without reconstructing one-dimensional signals. YOLO and RF-DETR object detectors were evaluated in combination with physiologically constrained post-processing that selects consistent detections across stimulus intensity levels. The best-performing configuration, based on an optimized YOLO26M detector, achieved 92.76% accuracy in identifying wave V presence, a mean absolute wave V latency error of 0.28 ms, and 94.3% agreement with expert-defined hearing thresholds within ±10 dB. Post-processing improved threshold agreement by enforcing consistency across intensity levels. These findings demonstrate the feasibility of extracting clinically relevant information directly from graphical ABR reports and suggest potential utility as decision support for audiologists. Further evaluation across clinical centers, recording devices, and report formats is required to establish generalizability.

Keywords: auditory brainstem response; hearing threshold detection; wave V latency; object detection; deep learning; computer vision

1. Introduction

The auditory brainstem response (ABR) is a short-latency auditory evoked potential that provides an objective functional measure of the early auditory pathway. Because ABR waveforms are recorded across stimulus intensities, they support estimation of hearing threshold and therefore play an important role in hearing screening, clinical diagnostics, and differential diagnosis. A typical ABR waveform contains several positive peaks, usually labeled I–VII [1], of which wave V is the most clinically relevant for threshold estimation. In practice, the hearing threshold is defined as the lowest stimulus intensity at which wave V can be reliably identified.

Although ABR can be represented as a digital time series, this representation does not fully reflect how the examination is typically interpreted in routine clinical practice. In many commercially used ABR systems, the output available to the clinician is a report image or PDF that contains patient information, acquisition parameters, and multiple plotted responses recorded at different stimulus intensity levels. These graphs are often visually complex, and waveforms obtained at neighboring intensities may appear in close proximity or partially overlap. Consequently, ABR interpretation in practice is not only a signal-processing task but also a visual pattern-recognition task applied to

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

medical images. Clinicians identify wave V by inspecting waveform morphology, relative timing, and characteristic relationships between peaks and troughs across plots.

This practical perspective is important because visual ABR reading is difficult and inherently subjective. Wave V determination, particularly near threshold, is affected by low signal quality, variable morphology, and uncertainty regarding the presence or absence of a true response [2,3]. As a result, interpretation is time-consuming and prone to both inter-observer and intra-observer variability. These limitations are especially relevant in large-scale neonatal hearing screening programs, where entire newborn populations are tested, while the number of highly experienced specialists remains limited. In such settings, scalable and reproducible automated support is of substantial clinical value [4–6].

Most previously reported automated ABR methods have been developed for digitally available waveform data. These approaches include classical feature-based classifiers, support vector machines, convolutional and recurrent neural networks, transformer-based models, time–frequency representations, and unsupervised clustering methods. They have shown promising performance in tasks such as response classification, wave annotation, latency estimation, and threshold detection [7–11]. However, most of these methods assume access to the underlying digital signal or to engineered representations derived from it. This assumption is not always compatible with real-world clinical workflows, in which the primary available artifact is often a device-generated image report rather than the raw sampled waveform.

Although plotted curves can, in principle, be converted back into digital signals, this introduces an additional preprocessing step that is not neutral. Recovering the waveform from a graph may alter its morphology, degrade local visual context, and introduce reconstruction errors that are difficult to quantify, particularly when plots overlap or contain superimposed visual elements. From the clinical perspective, a more direct and workflow-consistent alternative is therefore to analyze the report image itself. Despite this, image-based analysis remains relatively underexplored in the ABR literature, even though it more closely reflects the way ABR examinations are commonly reviewed in practice.

Motivated by this discrepancy between algorithmic assumptions and clinical reality, the present study proposes a deep-learning framework that operates directly on ABR graph images and formulates wave V identification as a computer vision task. Specifically, the method is designed to detect a characteristic visual configuration associated with wave V, enabling simultaneous determination of response presence and latency. Rather than relying on one-dimensional signal processing alone, the proposed approach is tailored to the graphical output of ABR devices and to the ambiguity inherent in visual interpretation. Against this background, the following section reviews previously reported automated methods for ABR analysis and situates the proposed framework within the broader literature.

2. Related Work

Previous studies on automated ABR analysis can be broadly grouped into several methodological categories, including classical non-machine-learning approaches, supervised machine-learning methods, deep-learning-based frameworks, multimodal or translational models, and image-based approaches.

Classical, non-machine-learning approaches have included both correlation-based and model-based methods operating directly on one-dimensional ABR waveforms. A resampling-assisted, correlation-based method for automatic ABR threshold estimation was introduced by Shaheen et al. [12]. In this approach, single-trial responses are repeatedly randomly partitioned to derive a distributional measure of response strength across stimulus levels. The method can be viewed as a classical correlation-based framework augmented with statistical resampling rather than as a purely statistical or machine-learning-based model. Developed and evaluated on murine ABR data, the method achieved 92% agreement with human-assessed hearing thresholds within a tolerance of ±10 dB. In a different line of work, Kamerer [13] proposed a model-based feature extraction method for neurodiagnostic ABR analysis, in which a Gaussian mixture model is used to represent waveform morphology. By

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

parameterizing individual ABR components as overlapping Gaussian functions, the method enables joint estimation of waveform features while accounting for interactions between adjacent waves. This design addresses a key limitation of peak-detection-based approaches, which largely mimic expert visual inspection and may therefore inherit its characteristic errors. However, the method was developed for wave identification and morphological characterization rather than for threshold estimation and is therefore not intended for hearing-threshold detection in its current form.

A classical supervised machine-learning approach to ABR threshold detection was proposed by Acır et al. [14], in which one-dimensional ABR waveforms were represented using handcrafted features, including raw amplitudes, DCT coefficients, and DWT approximation coefficients, followed by adaptive feature selection and binary SVM classification. Although the reported classification accuracy exceeded 97%, this result was obtained on a relatively small dataset comprising only 648 records, of which 468 were used for training. Moreover, the study did not address latency estimation.

More recent work has increasingly adopted data-driven and deep-learning-based approaches. An interpretable deep-learning framework for automated ABR detection was proposed by Liu et al. [15], in which one-dimensional ABR waveforms were transformed into two-dimensional time-frequency representations using continuous wavelet transform and subsequently classified by a Vision Transformer-based model (CWT-ViT). The CWT-ViT model achieved an accuracy of 91.83% on the held-out test set and 93.54% on an independent external dataset. However, the study focused on binary ABR response detection from signal-derived time–frequency representations rather than direct interpretation of graphical clinical reports or hearing-threshold estimation. In another study, Liu et al. [16] introduced a lightweight deep learning model for automated recognition of ABR wave latencies. The proposed Patch-DAT architecture combines a patch-based representation of the waveform with a first-derivative-guided dual-attention mechanism, allowing efficient modelling of both local temporal patterns and global dependencies. When evaluated on large multi-center datasets, the model achieved high latency-recognition accuracy and showed strong generalization across clinical settings. Liang et al. [17] proposed a feature-based deep learning framework for automatic recognition of ABR waveforms in participants with varying ages and hearing levels. In this approach, time- and frequency-domain features extracted from ABR signals were combined with demographic and clinical variables, such as age, sex, pure-tone thresholds, stimulus intensity, and raw signal sequences, to form the input representation. The method was evaluated both on the full dataset and on subgroups defined by age and hearing status and achieved a weighted mean accuracy exceeding 95%.

Hybrid machine-learning approaches combining conventional statistical descriptors with learned representations have also been investigated. McKearney et al. [18] proposed a supervised machine-learning framework for ABR detection in which multiple candidate models were evaluated and a stacked ensemble was ultimately selected as the best-performing approach. The final model combined a random forest operating on statistically derived features with a CNN-LSTM operating on waveform-based temporal inputs. Their outputs were then fused through a logistic regression meta-classifier. The method can therefore be characterized as a hybrid stacked-ensemble machine-learning approach for binary ABR detection. A key limitation of this approach is its dependence on the quality of the training annotations, which is a general drawback shared by supervised learning methods.

Beyond direct threshold-detection studies in humans, machine learning has also been applied to ABR data from translational animal models. Such studies are particularly valuable when clinical datasets are limited, as animal ABRs remain sufficiently comparable in morphology to support methodological development despite species-specific differences in hearing range. Ceriani et al. [19] proposed machine-learning-based methods for the early prediction of progressive hearing loss from mouse ABR time series, enabling both threshold detection and the assessment of progressive changes. These findings support the broader translational potential of ABR-based ML approaches for future human diagnostic applications. A related but methodologically distinct example was presented by Erra et al. [20], who addressed the problem of latency variability across ABR waveforms recorded at

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

different frequencies and stimulus intensities. To improve comparability in the time domain, they proposed the optional use of time warping to align characteristic waveform features. This procedure was incorporated into an unsupervised threshold estimation approach based on functional principal component analysis (FPCA) and k-means clustering. The method reduces waveform dimensionality and partitions responses into groups corresponding to above- and below-threshold activity without relying on predefined labels. In a broader multimodal context, Manta et al. [21] developed a machine-learning framework for classifying tinnitus-related distress using auditory evoked potentials and clinical data. Their approach combined features derived from ABR and AMLR waveforms, including wavelet-scattering coefficients in the time-frequency domain, with selected clinical characteristics. Among the evaluated feature sets, wavelet-scattered AMLR signals were the most informative, and their combination with 15 LASSO-selected clinical features yielded the best performance, with an AUC of 92.53%, sensitivity of 84.84%, and specificity of 83.04%.

A less common line of research treats ABR data as images rather than as directly processed waveforms. Ma et al. [22] presented a VGG16-based approach for classifying preprocessed ABR graph images into normal-hearing and hearing-loss categories. Their method relied on image standardization procedures designed to harmonize ABR plots acquired from different clinical devices, including axis normalization, grayscale conversion, removal of waveform markers, and image-size normalization. However, the objective of that study was limited to binary hearing-loss classification rather than ABR threshold detection or latency estimation.

A related image-based approach to ABR threshold estimation was presented by Nowakowski and Czyżewski [23], who investigated the use of two-dimensional discrete wavelet transforms (2D-DWT), template matching, and support vector machine classification. Their study systematically evaluated 106 wavelet functions and six region-of-interest sizes, comparing intensity-specific and global waveform templates. The best-performing configuration achieved 89.94% agreement with expert-defined ABR thresholds within a tolerance of ±10 dB. The study was conducted using an earlier, smaller version of the clinical dataset employed in the present work. Unlike the wavelet-based feature extraction and template-matching approach, the present framework employs deep learning object detection combined with physiologically constrained post-processing, enabling both hearing-threshold estimation and wave V latency determination.

Most methods reported in the literature assume direct access to digitally represented ABR waveforms, which are well-suited for signal analysis and automated processing. In clinical practice, however, ABR outputs are often stored or presented as report-like graphical records that include patient information, examination parameters, and waveform plots across stimulus intensity levels. In such cases, conventional digital signal processing requires an additional waveform digitization step, which may introduce reconstruction errors and associated uncertainty. To address this practical limitation, the present study adopts a less commonly explored image-based approach using computer-vision methods.

The principal contributions of this study can be summarized as follows:

1. the direct analysis of device-generated ABR reports and graph images,
2. a unified framework enabling simultaneous wave V detection, latency estimation, and hearing-threshold determination,
3. validation under conditions representative of routine clinical practice.

The remainder of this paper is organized as follows. We first describe the dataset and the preprocessing procedures used to extract intensity-specific graph images from source reports. We then present the proposed detection and post-processing framework, followed by experimental evaluation of several deep learning architectures. Finally, we discuss the clinical implications of the results and the potential of computer-vision-based ABR analysis for routine practice.

3. Materials and Methods

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

The proposed method follows a general workflow consisting of five main stages, all implemented in Python scripts:

1. Input data preprocessing (preparation of the training, validation, and test sets according to the requirements of the machine learning model used),
2. Model training,
3. Inference,
4. Post-processing,
5. Final report generation.

Preprocessing and model training are preparatory steps that can be performed once and repeated only occasionally. Retraining may be required when a deterioration in prediction quality is observed, for example, because of a change in the data source. The individual stages and the dataset used are described in detail in the following subsection.

3.1. Dataset

To support supervised machine learning, the graphical data were annotated by experts from the Institute. The annotations were stored on a separate graphical layer and consisted of a line connecting the peaks of wave V, which allowed both the presence of wave V and its latency to be determined.

The study was based on clinical auditory brainstem response (ABR) data provided by the Institute of Physiology and Pathology of Hearing in Warsaw, Poland. The recordings were acquired using the Vivosonic Integrity V500 system and stored as PDF reports containing ABR waveforms recorded at different acoustic stimulus intensity levels.

The original dataset comprised 16,757 PDF reports. The age and sex distributions of the study population are presented in Figure 1. Children younger than 10 years accounted for approximately 55% of the patients, while adults aged 40–70 years represented approximately 30%. The remaining age groups accounted for 15%. The proportion of male patients was 54%.

To prevent data leakage, the original PDF reports were partitioned into a development subset (15,105 PDF reports) and an independent test subset (1652 PDF reports) before extracting individual ABR graph images. The partitioning procedure incorporated patient-level grouping and stratification by ear and acoustic stimulus type. All reports belonging to the same patient were assigned exclusively to one subset, ensuring that no patient contributed data to both development and testing.

The development subset was used for model training, validation, hyperparameter optimization, and post-processing parameter selection. It comprised 42,464 training graphs and 10,616 validation graphs. The independent test subset was reserved exclusively for final evaluation and was not used for model development or parameter tuning.

Because individual PDF reports contained different numbers of stimulus-specific ABR graphs, the proportions of extracted graph images did not necessarily correspond to the original 90:10 partition of the source reports.

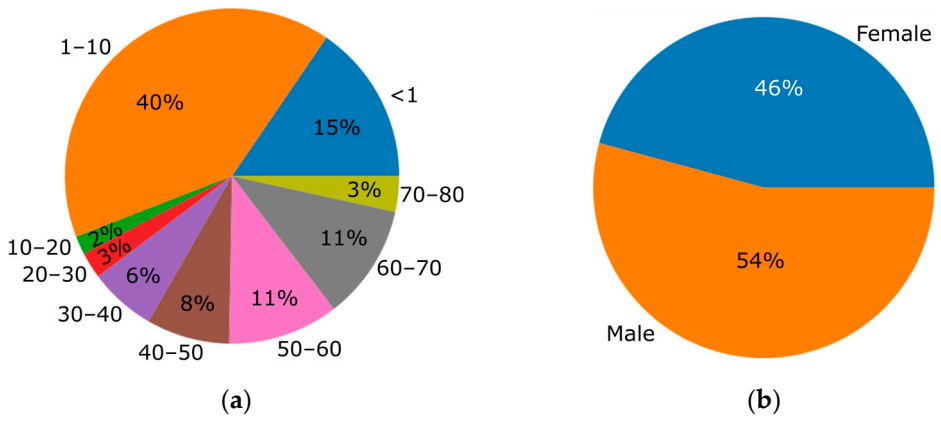


Figure 1. Distribution of the study population by (**a**) age and (**b**) sex.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

The reference annotations were prepared at the collaborating clinical institute by experienced audiologists as part of the manual interpretation of ABR reports. According to information provided by the collaborating specialists, multiple audiologists participated in the annotation process, although the exact number of annotators and the allocation of reports among them could not be established.

The annotation procedure consisted of identifying wave V peaks across the recorded stimulus intensity levels and connecting their positions with straight line segments. To facilitate the identification of ambiguous responses, particularly near the hearing threshold, specialists could manipulate the graphical representation by horizontally compressing and vertically stretching the waveforms, thereby emphasizing morphological features relevant to wave V identification.

The resulting annotations were stored in a separate graphical layer within the original PDF reports. These annotated reports were subsequently provided as the source dataset for the present study.

Detailed documentation of the annotation protocol, independent review procedures, and inter-rater agreement statistics was not available to the authors. Consequently, the expert-provided annotations were treated as the reference standard for model development and evaluation, while the potential uncertainty associated with individual expert judgments was acknowledged as a limitation of the study.

An earlier subset of this clinical dataset was used in a previous study investigating wavelet-based ABR threshold estimation [23]. The present study extends the dataset and investigates a distinct deep-learning-based detection and post-processing framework.

3.2. Preprocessing

3.2.1. Extraction of Intensity-Specific ABR Graph Images

The ABR graphs were extracted from the original PDF reports as raster images with a variable height and a width of at least 890 pixels. The original time axis typically extended from approximately −0.5 to 24.5 ms. An example is presented in Figure 2, showing the original waveform plots and the corresponding expert annotations stored in a separate graphical layer.

During preprocessing, each image was cropped to retain the time interval from 0 to 20 ms and subsequently resized to 640 × 640 pixels. This normalization established a uniform horizontal scale of 32 pixels per millisecond, enabling direct conversion between image coordinates and wave V latencies.

The normalized images contained multiple waveform groups corresponding to different acoustic stimulus intensity levels. Each group was extracted separately and vertically centered within a 640 × 640-pixel image while preserving the normalized horizontal time scale. Examples of the resulting intensity-specific images are presented in Figure 3.

The extraction procedure removed graphical elements unnecessary for waveform analysis, including the grid and waveform markers. Expert annotations stored in a separate PDF layer were excluded from the model input images. The resulting images therefore contained the ABR waveform information required for subsequent object detection without the graphical reference annotations.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

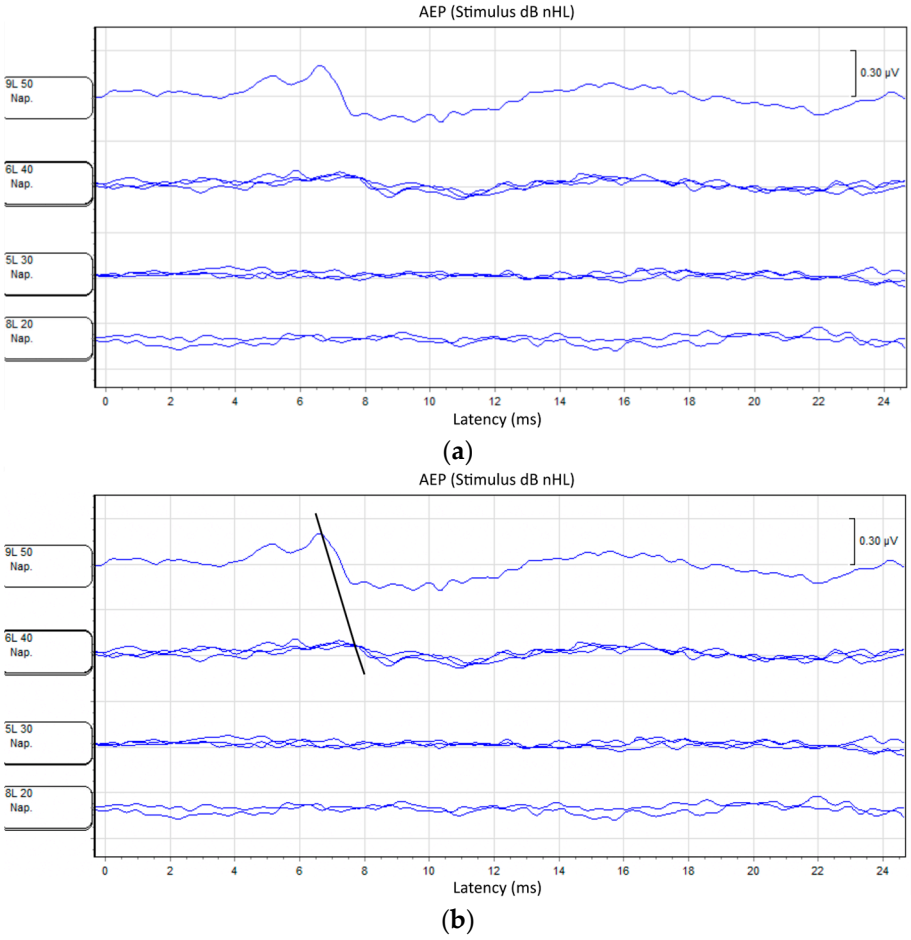


Figure 2. Example of an ABR report containing responses recorded at different stimulus intensity levels: (a) original waveforms without expert annotations; (b) the same waveforms with the expert annotation layer enabled. The annotation line connects the identified wave V peaks and was used to extract reference peak coordinates and corresponding latencies. The annotation layer was excluded from the images supplied to the detection models.

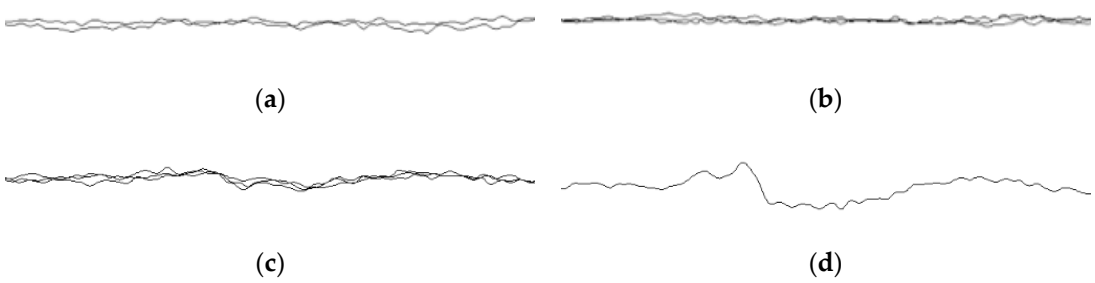


Figure 3. Examples of intensity-specific ABR graph images extracted from the report shown in Figure 2: (a) 20 dB, (b) 30 dB, (c) 40 dB, and (d) 50 dB.

The resulting dataset comprised 53,080 graphs in total, including 26,385 with a detectable wave V. The distribution of stimulus types was as follows: 0.5 kHz – 14%, 1 kHz – 37%, 2 kHz – 6%, 4 kHz – 5%, and click stimuli – 37%.

3.2.2. Generation of Reference Labels from Expert Annotations

The original PDF reports were annotated by experienced audiologists, who identified wave V peaks across the recorded stimulus intensity levels. The identified peaks were connected by straight line segments stored in a separate graphical layer of the PDF document (Figure 2b).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

During preprocessing, the coordinates of the intersections between the expert-drawn annotation lines and the corresponding waveform plots were extracted automatically. These coordinates were treated as reference wave V peak positions and were used to determine the corresponding reference latencies and generate object detection labels.

Two annotation classes were defined: "wave", representing the presence of an identifiable wave V, and "noWave", representing the absence of an identifiable wave V within the expected latency interval.

For the "wave" class, the bounding box was positioned relative to the reference wave V peak. Its left boundary was placed 32 pixels (1 ms) before the peak to include the preceding ascending waveform slope, while its upper boundary was positioned 64 pixels above the peak. The bounding-box width was fixed at 256 pixels (8 ms), whereas its height was adjusted according to the vertical extent of the waveform. This configuration encompassed the characteristic wave V peak and the subsequent negative deflection (Figure 4). These morphological features are also considered by audiologists during routine visual interpretation of ABR waveforms [24–26], providing a clinically motivated basis for the selected bounding-box geometry. The bounding-box parameters were determined empirically by evaluating the impact of different configurations on model training and validation performance.

For the "noWave" class, where no reference wave V peak was identified, a fixed bounding box of 353 × 60 pixels was used. The box was horizontally positioned to cover approximately the 3–14 ms latency interval and vertically centered in the normalized 640 × 640-pixel image. Unlike the "wave" class, the "noWave" bounding box did not represent a localized physiological response, but rather a predefined region of interest in which the characteristic wave V morphology was considered absent.

The use of these two classes enabled the object detection framework to jointly represent the presence and spatial localization of wave V and its absence within a physiologically motivated region of interest. The fixed relative geometry of the "wave" bounding box also allowed the wave V latency to be recovered from its predicted horizontal position.

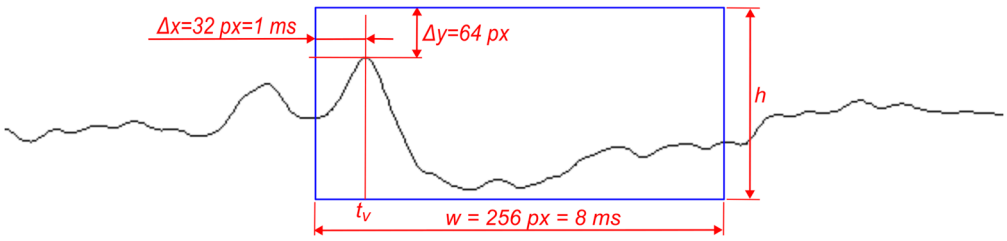


Figure 4. Definition of the bounding-box annotations for the "wave" detection class.

3.3. Model Training

3.3.1. Model Architectures

The proposed framework formulates wave V identification as an object detection task, in which characteristic waveform patterns are localized and classified in intensity-specific ABR graph images. Given the relatively small size and subtle morphology of the target patterns, two families of deep learning object detectors were investigated: convolutional neural network-based YOLO models and transformer-based RF-DETR models.

The YOLO family was selected for its one-stage detection architecture, which performs object localization and classification within a unified inference process. In contrast, RF-DETR builds upon the transformer-based object detection paradigm introduced by Carion et al. [27], enabling the evaluation of an alternative approach to learning spatial relationships and visual representations.

The experimental evaluation covered several variants of the YOLO11, YOLO26, and RF-DETR architectures, including models with different capacities, training configurations, and hyperparameter optimization strategies. Additional experiments investigated the effects of pretrained weights and

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

input resolution on RF-DETR performance. The latter investigation was motivated by the transformer-based architecture of RF-DETR and its image representation and processing requirements.

Table 4 presents a representative selection of the evaluated model configurations rather than an exhaustive list of all experiments conducted. This selection enables comparison of the effects of model architecture, training strategy, and, specifically for RF-DETR, input resolution and pretraining on the final hearing-threshold estimation performance.

All models were trained to distinguish between two annotation classes, "wave" and "noWave", using the standardized 640 × 640-pixel ABR graph images and bounding-box annotations described in Section 2.2. The predicted detections were subsequently processed using the same physiological constraint-based post-processing framework described in Section 3.5.

3.3.2. Training Configuration

The YOLO26M model was trained using the Ultralytics framework (version 8.4.9) with PyTorch 2.8.0 and CUDA 12.8 on an NVIDIA H200 NVL GPU. The training and validation datasets comprised 42,464 and 10,616 annotated images, respectively. All images were processed at a resolution of 640 × 640 pixels, and the detection head was configured to recognize two classes: "wave" and "noWave".

Training was initialized using pretrained YOLO26M weights. An initial baseline experiment was conducted for 300 epochs with a batch size of 16 and the default Ultralytics training configuration. Automatic mixed-precision training was enabled, and the optimizer was selected automatically by the framework, resulting in the use of MuSGD with an initial learning rate of 0.01 and momentum of 0.9.

Following hyperparameter optimization, the final YOLO26M model was trained with a batch size of 128 and a maximum of 300 epochs. Early stopping was applied with a patience of 50 epochs. Training terminated after 102 epochs, with the best checkpoint obtained at epoch 52.

Model selection was based on validation performance. The selected checkpoint achieved an mAP@0.50 of 0.957 and an mAP@0.50:0.95 of 0.894 on the validation dataset. The independent test dataset was not used during training, hyperparameter optimization, or checkpoint selection.

The RF-DETR models were trained using the RF-DETR framework with PyTorch and CUDA acceleration. The training and validation datasets comprised 42,464 and 10,616 annotated images, respectively, and contained the same two detection classes, "wave" and "noWave", as the YOLO experiments.

Training employed separate learning rates for the main network and the image encoder, together with weight decay regularization. The experiments employed automatic mixed precision (AMP), gradient accumulation, and an exponential moving average (EMA) of model parameters to support training stability. Early stopping was enabled based on validation performance.

Several RF-DETR model configurations were investigated, including variants initialized with weights pretrained on the Microsoft COCO dataset [28] and variants trained without such pretraining. Unlike the YOLO experiments, which used a fixed input resolution of 640 × 640 pixels, the RF-DETR experiments also explored different input resolutions to assess their influence on detection performance. This investigation was motivated by the transformer's image representation and processing characteristics.

The training configurations were evaluated using bounding-box detection metrics, including mAP@0.50 and mAP@0.50:0.95. The resulting models were subsequently assessed within the complete ABR interpretation framework using the hearing-threshold estimation procedure described in Section 3.5.

3.3.3. Hyperparameter Optimization

Hyperparameter optimization was performed using the Optuna framework [29] with the Tree-structured Parzen Estimator (TPE) sampling algorithm. For the YOLO26M model, the optimization comprised 30 sequential trials, each allowing up to 60 training epochs with a batch size of 128 and an input image resolution of 640 × 640 pixels. Early stopping was applied with a patience of 10 epochs.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

The search space comprised 16 parameters related to learning-rate scheduling, regularization, loss-component weighting, and data augmentation. These included the initial and final learning-rate factors, momentum, weight decay, warm-up duration, bounding-box regression loss weight, classification loss weight, distribution focal loss weight, label smoothing, HSV-based color augmentation, image scaling, horizontal flipping, mosaic augmentation, and mixup.

The optimization objective was to minimize the negative mean average precision averaged over intersection-over-union thresholds from 0.50 to 0.95:

$$f(\theta) = -\mathrm{mAP}_{50:95}(\theta) \quad (1)$$

where θ denotes the vector of candidate hyperparameters. The best configuration was obtained in trial 17, with an objective value of −0.8947, corresponding to a validation mAP@0.50:0.95 of approximately 0.895.

The Ultralytics automatic optimizer selection mechanism was enabled throughout the experiments. Consequently, the candidate initial learning rate and momentum values supplied by Optuna were overridden by the framework. The effective optimizer was MuSGD with an initial learning rate of 0.01 and momentum of 0.9.

The hyperparameters selected in the best trial were subsequently used for final model training, with the maximum number of epochs increased to 300 and early-stopping patience increased to 50. The independent test dataset was not used during hyperparameter optimization or checkpoint selection.

Hyperparameter optimization of RF-DETR was performed using Optuna with the Tree-structured Parzen Estimator (TPE) algorithm and a fixed random seed of 42. Two independent optimization studies were conducted, each comprising 50 scheduled trials with a maximum of 30 training epochs per trial.

Both studies employed the same two-class detection task but differed in their optimization objectives. The first study maximized the class-specific mAP@0.50:0.95 for "wave", prioritizing accurate localization of the wave V morphology. The second study maximized the overall mAP@0.50 across both detection classes, emphasizing general detection performance.

The hyperparameter search included the requested input resolution (616, 672, or 728 pixels), learning rate, encoder learning rate, weight decay, batch size, and gradient accumulation steps. The effective image resolution was additionally influenced by the internal resizing and multi-scale processing mechanisms of the RF-DETR implementation.

Each trial trained an RF-DETR model with automatic mixed precision, exponential moving average of model parameters, and early stopping enabled. The objective values were retrieved directly from the validation metrics generated by the RF-DETR training framework.

The wave-focused study selected trial 40, achieving a validation mAP@0.50:0.95 of 0.8097 for the "wave" class and 0.8586 across both classes. The overall-detection study selected trial 24, achieving an overall validation mAP@0.50 of 0.9499 and mAP@0.50:0.95 of 0.8611. These results were used to compare the effects of the two optimization objectives on detection performance.

Preliminary experiments involved YOLO11 models of three sizes (Small, Medium, and Large), trained using the same two-class annotation scheme as the subsequent YOLO26 experiments. Several training configurations were evaluated, including baseline settings and hyperparameters optimized using Optuna. These experiments served to establish a performance reference and assess the influence of model capacity on ABR waveform detection. Since increasing model size did not consistently improve performance, subsequent development focused on medium-sized architectures. Representative YOLO11 results are retained in Table 4 for comparison with YOLO26 and RF-DETR.

3.3.4. Preliminary Experiments

Preliminary experiments were conducted to investigate the influence of stimulus type on wave V latency distributions and detection performance. Figure 5 presents the estimated latency distributions for individual stimulus types across different stimulus intensity levels. The distributions varied

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

according to stimulus type, indicating that a single latency window may not be equally representative of all recording conditions. The corresponding latency intervals are summarized in Table 1.

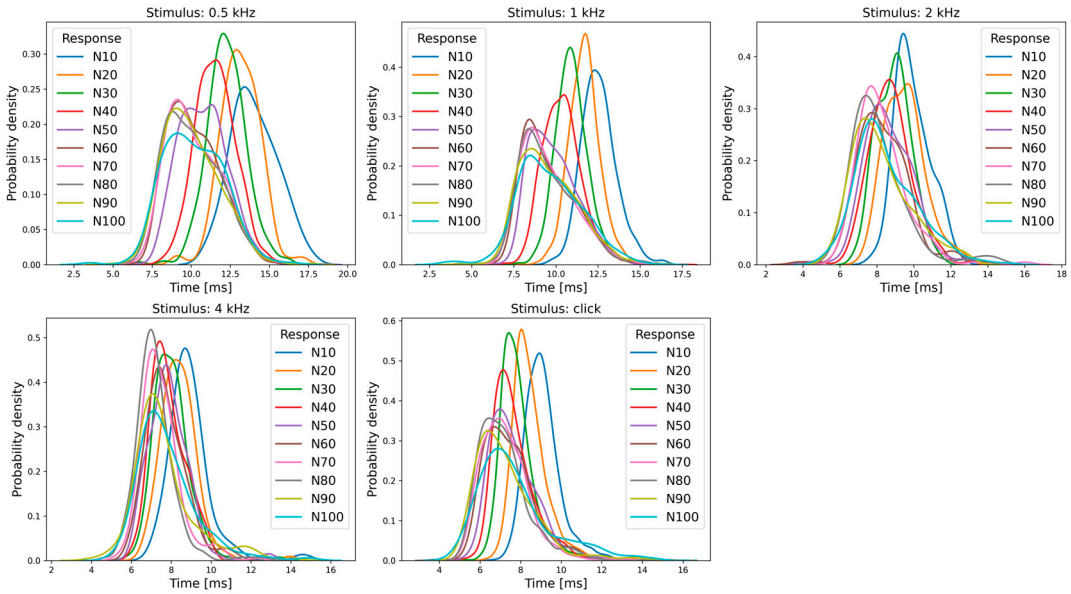


Figure 5. Latency distributions for individual types of stimulation.

Table 1. Latency window depending on the type of stimulus.

Stimulus	Latency window
0.5 kHz	6–17 ms
1 kHz	6–15 ms
2 kHz	5–13 ms
4 kHz	5–11 ms
Click	4.5–12 ms

To assess the influence of stimulus-dependent latency constraints, two inference configurations were initially compared using data from the development subset: a fixed latency window of 5–16 ms and stimulus-specific windows defined in Table 1. As shown in Table 2, the fixed-window approach achieved slightly higher recall, F1-score and accuracy, whereas stimulus-specific windows yielded marginally higher precision, and a lower mean absolute latency error (0.29 ms versus 0.32 ms). Given these relatively small differences, a common latency window was preferred for subsequent development. The initial 5–16 ms interval was subsequently refined during post-processing optimization, resulting in a final interval of 3–14 ms, which provided better overall hearing-threshold estimation performance.

Table 2. Results for fixed and variable latency windows. Total images: 5019; wave: 2414; noWave: 2605.

Metric	Results for a fixed latency window	Results for a variable latency window
TP	2125	2088
FP	151	130
TN	2454	2475
FN	289	326
Precision	93.37%	94.14%
Recall	88.03%	86.50%
F1-score	90.62%	90.16%
Accuracy	91.23%	90.91%
Latency MAE	0.32 ms	0.29 ms

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

Additional experiments investigated the influence of stimulus type on model training. A YOLO11M detector was trained and evaluated using different stimulus-specific subsets, as summarized in Table 3. The resulting precision, recall, and F1 scores were broadly comparable across the investigated configurations. These findings suggest that the detection approach was not strongly dependent on the composition of the evaluated stimulus subsets under the experimental conditions considered. However, they do not establish statistical independence between waveform morphology and stimulus type.

Table 3. Training results of the YOLO11 M model.

Model:	YOLO11 M						
Number of epochs:	200						
Number of classes:	2 (wave, noWave)						
Type of stimulus	Number of images	wave	noWave	Ratio wave/noWave	Precision	Recall	F1 score
Click	1425	847	578	1.465	0.91	0.90	0.91
1 kHz	1569	820	749	1.095	0.88	0.88	0.88
0.5+1+2+4 [kHz]	2771	1282	1489	0.861	0.89	0.89	0.89
all	4196	2129	2067	1.03	0.90	0.90	0.90

3.4. Detection of Wave V in Each Group

During inference, the trained object detection model processes each intensity-specific ABR graph separately. For each image, the model produces a set of candidate detections, each characterized by a class label ("wave" or "noWave"), a bounding box, and an associated confidence score.

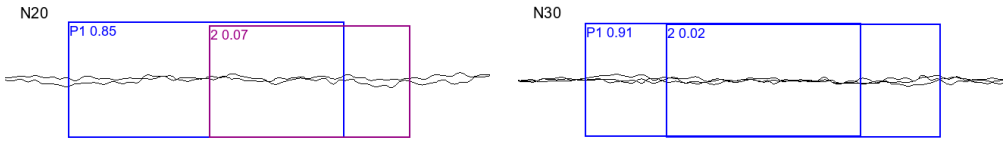
A low confidence threshold of 0.01 was adopted to retain potentially relevant detections, including those associated with ambiguous waveform morphology near the hearing threshold. Consequently, multiple candidate detections may be retained for a single intensity level. Examples are presented in Figure 6, where the bounding boxes illustrate the predicted classes and their corresponding confidence scores.

For detections classified as "wave", the horizontal coordinate of the bounding box was used to estimate the wave V latency. As described in Section 3.2.2, the reference bounding-box geometry places its left boundary 32 pixels (1 ms) before the wave V peak. Therefore, for a normalized 640 × 640-pixel image representing a 0–20 ms interval, the estimated latency was calculated as

$$t_V = \frac{x_{\text{box}} + 32}{32} \text{ ms} \quad (2)$$

where x_{box} denotes the horizontal coordinate of the predicted bounding box's left boundary. Detections classified as "noWave" represent the absence of an identifiable wave V and do not provide a wave V latency estimate.

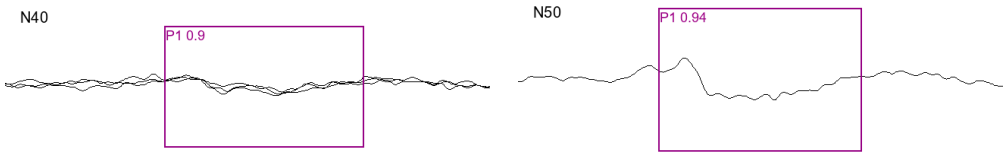
At this stage, detections are treated as individual candidates rather than final decisions. Their confidence scores, class assignments, and, where applicable, estimated latencies are subsequently evaluated jointly across stimulus intensity levels using the physiological constraints and combination-selection procedure described in Section 3.5.



(**a**) (**b**)

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1



(**c**) (**d**)

Figure 6. Examples of candidate object detections in intensity-specific ABR graph images: (a) 20 dB, (b) 30 dB, (c) 40 dB, and (d) 50 dB. Bounding boxes corresponding to the "noWave" and "wave" classes are shown in blue and purple, respectively. Each detection is labeled with an index and its confidence score. Multiple candidate detections may be retained for the same intensity level and are subsequently evaluated during post-processing.

3.5. Post-Processing

As a result of the inference mentioned in the previous section, each stimulus intensity level can be assigned a set of detections. The most important element of post-processing is identifying the correct detection in each of these sets. The method used is presented in the next section.

3.5.1. Detection Combination and Selection

Let

$$\mathcal{N} = \{N^{(i)}\}_{i=1}^{L} \tag{3}$$

denote the ordered set of stimulus intensity levels considered in the analysis.

For example, for the image shown in Figure 2

$$N = \{N20, N30, N40, N50\}, \qquad L = 4. \tag{4}$$

For each intensity level $N^{(i)}$, the AI model produces a set of candidate detections

$$D_{N^{(i)}} = \{d_{i,1}, d_{i,2}, \dots, d_{i,m_i}\}, \tag{5}$$

where $d_{i,j}$ denotes the j-th detection for level $N^{(i)}$, and $m_i = | D_{N^{(i)}} |$ is the number of detections at that level (Figure 6). For each stimulus intensity level, at least one candidate detection belonging to either the "wave" or "noWave" class was retained, ensuring that the corresponding detection set was never empty.

All possible detection combinations are defined as the Cartesian product

$$C = \prod_{i=1}^{L} D_{N^{(i)}}. \tag{6}$$

Each element of C is a tuple

$$C_k = (d_{1,j_1}, d_{2,j_2}, \dots, d_{L,j_L}), \tag{7}$$

where $d_{i,j_i} \in D_{N^{(i)}}$.

The total number of combinations is therefore

$$K = | C | = \prod_{i=1}^{L} | D_{N^{(i)}} |. \tag{8}$$

Each detection $d_{i,j}$ is associated with a class label $y_{i,j} \in \{0,1\}$, confidence score $p_{i,j}$, and latency $Lat_{i,j}$. Here $y_{i,j} = 1$ denotes the presence of wave V and $y_{i,j} = 0$ corresponds to the absence of the wave.

Candidate combinations are filtered using physiological constraints. First, the latency of detected waves must lie within the interval

$$\text{MIN_LATENCY} \le Lat_{i,j_i} \le \text{MAX_LATENCY}. \tag{9}$$

Second, detections must be monotonic with respect to stimulus intensity:

$$y_{i,j_i} \le y_{i+1,j_{i+1}}, i = 1, \dots, L-1. \tag{10}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

The stimulus intensity levels were ordered in ascending order. Since the class labels were encoded as 0 for "noWave" and 1 for "wave", the monotonicity constraint prevented a previously identified wave V response from disappearing at higher stimulus intensities.

Finally, the latency differences between adjacent stimulus intensity levels were constrained to reflect the expected reduction in wave V latency with increasing stimulus intensity, while allowing limited deviations due to waveform variability and measurement uncertainty. For two adjacent intensity levels at which wave V was detected, the constraint was defined as

$$-\text{LAT_MAX_DELAY} \leq (Lat_{i,j_i} - Lat_{i+1,j_{i+1}}) \leq \text{LAT_MAX_HASTE} \quad (11)$$

Here, LAT_MAX_DELAY denotes the maximum permitted increase in latency (1.25 ms), whereas LAT_MAX_HASTE denotes the maximum permitted decrease in latency (3.00 ms) as stimulus intensity increases. The constraint was applied only when both selected detections belonged to the "wave" class.

For each remaining valid combination C_k , the combination score $S(C_k)$ is defined as

$$S(C_k) = \prod_{i=1}^{L} w(y_{i,j_i})\, p_{i,j_i}, \quad (12)$$

where p_{i,j_i} denotes the confidence score of the selected detection and $w(\cdot)$ is a class-dependent weight

$$w(y_{i,j_i}) = \begin{cases} 1 & y_{i,j_i} = 1 \\ \text{CLASS_BAL} & y_{i,j_i} = 0 \end{cases} \quad (13)$$

The score was used exclusively for ranking candidate combinations and was not interpreted as a calibrated probability.

The optimal detection configuration is obtained by maximizing the combination score

$$C^* = \arg \max_{C_k \in \mathcal{C}_{\text{valid}}} S(C_k). \quad (14)$$

Following selection of the highest-scoring physiologically admissible combination, the hearing threshold was estimated as the lowest stimulus intensity at which wave V was identified. For all intensity levels classified as "wave", the corresponding latency estimates were retained. Thus, the selected combination simultaneously provided the estimated hearing threshold and the intensity-specific wave V latencies.

If the selected combination contained only "noWave" detections, the result was classified as no detectable ABR response within the tested stimulus intensity range. In such cases, no numerical hearing threshold was assigned.

3.5.2. Algorithm Optimization

The algorithm was optimized by adjusting the following parameters:

1. the confidence threshold – value 0.01
2. the margins MIN_{L}ATENCY= 3 ms and MAX_{L}ATENCY= 14 ms
3. the allowable latency difference between adjacent intensity levels
4. LAT_MAX_DELAY = 1.25 ms and LAT_MAX_HASTE = 3 ms
5. the weight assigned to the „noWave" class, CLASS_BAL = 0.5

The latency interval was initially defined as 5–16 ms based on the empirical distributions of wave V latencies (Figure 5 and Table 1). Subsequent optimization of the post-processing algorithm, conducted exclusively on the development dataset, demonstrated improved hearing-threshold estimation performance with a shifted interval of 3–14 ms. This interval was therefore adopted in the final configuration.

3.6. Algorithm Evaluation

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

The proposed framework was evaluated at two complementary levels: wave V detection in individual intensity-specific ABR images and hearing-threshold estimation at the report level. The evaluation included representative configurations of the YOLO11 [30,31], YOLO26 [32] and RF-DETR [33–36] model families, including baseline and hyperparameter-optimized variants.

Wave V detection performance was assessed using the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), with the presence of wave V treated as the positive class. Accuracy, precision, recall, and F1-score were calculated from the corresponding confusion matrices. Wave V latency estimation was evaluated using the mean absolute error (MAE), calculated exclusively for true-positive detections, i.e., cases in which both the expert reference and the model indicated the presence of wave V. The absolute differences between the reference and predicted horizontal peak coordinates were converted to milliseconds using the normalized image scale of 32 pixels per millisecond and then averaged.

Hearing-threshold estimation was assessed at the report level by comparing the estimated threshold with the expert-defined reference threshold. Two metrics were used: OK, denoting exact agreement with the reference threshold (0 dB difference), and OK_10, denoting agreement within a tolerance of ±10 dB.

As previously stated, the final evaluation dataset comprised 1652 PDF reports. Of these, 29 reports could not be processed successfully, leaving 1623 reports for hearing-threshold evaluation. These reports yielded 6116 intensity-specific ABR images, including 3184 annotated as containing wave V and 2932 annotated as not containing an identifiable wave V.

Table 4 summarizes the performance of selected model configurations within the complete ABR interpretation framework. The table includes different model capacities, hyperparameter-optimized variants, and RF-DETR configurations differing in pretraining strategy and input resolution. The presented models constitute a representative subset of the configurations investigated during development.

Table 4. Performance of selected YOLO and RF-DETR configurations for wave V detection, latency estimation, and hearing-threshold determination. Detection metrics were calculated at the image level, whereas hearing-threshold agreement (OK and OK_10) was evaluated at the report level.

Model	TN	TP	FN	FP	ACC	P	R	F1	Latency MAE	OK	OK_10
YOLO26M_Optuna	2767	2906	278	165	92.76%	94.63%	91.27%	92.92%	0.28 ms	76.50%	94.30%
YOLO11 M Optuna	2735	2943	241	197	92.84%	93.73%	92.43%	93.07%	0.27 ms	76.80%	94.00%
YOLO11 M	2752	2892	292	180	92.28%	94.14%	90.83%	92.46%	0.30 ms	75.00%	93.80%
YOLO11 L Optuna	2768	2879	305	164	92.33%	94.61%	90.42%	92.47%	0.28 ms	74.90%	93.70%
YOLO11 S	2756	2889	295	176	92.30%	94.26%	90.73%	92.46%	0.30 ms	75.50%	93.50%
YOLO11 L	2747	2883	301	185	92.05%	93.97%	90.55%	92.23%	0.30 ms	74.60%	93.30%
RF-DETR Large nopretrain_640	2718	2904	280	214	91.92%	93.14%	91.21%	92.16%	0.29 ms	74.40%	93.30%
RF-DETR Base	2732	2894	290	200	91.99%	93.54%	90.89%	92.19%	0.29 ms	74.70%	93.20%
RF-DETR Large nopretrain_704	2662	2941	243	270	91.61%	91.59%	92.37%	91.98%	0.30 ms	73.30%	93.00%
RF-DETR Large pretrain_640	2724	2878	306	208	91.60%	93.26%	90.39%	91.80%	0.30 ms	73.30%	92.90%
RF-DETR Medium nopretrain_704	2648	2946	238	284	91.47%	91.21%	92.53%	91.86%	0.30 ms	73.00%	92.50%
RF-DETR_Optuna	2596	2999	185	336	91.48%	89.93%	94.19%	92.01%	0.30 ms	73.30%	92.40%
YOLO26M	2766	2823	361	166	91.38%	94.45%	88.66%	91.46%	0.30 ms	73.00%	92.20%

Explanation of model names and abbreviations:

- *_Optuna – a model optimized using the Optuna library.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

- *_nopretrain_640 – a model trained without pretraining and with the input image size set to 640x640.
- *_704 – model trained with the input image size set to 704x704.
- *pretrain* - model pretrained on the COCO dataset.
- OK – classification accuracy with a margin of error of 0 dB relative to the ground truth (GT). In this case, the GT was taken as the expert-provided reference annotations.
- OK_10 classification accuracy with a margin of error of 10 dB relative to GT.

3.7. Algorithm Performance Analysis

To assess the contribution of the proposed post-processing algorithm, an ablation experiment was conducted using the optimized YOLO26M model (YOLO26M_Optuna). The model was evaluated with post-processing enabled and disabled, while maintaining the same evaluation dataset and detection model.

As shown in Table 5, post-processing improved wave V detection performance. The number of false negatives decreased from 375 to 278 (a relative reduction of 25.87%), while false positives decreased from 178 to 165 (7.30%). Consequently, classification accuracy increased from 90.96% to 92.76%, recall from 88.22% to 91.27%, and F1-score from 91.04% to 92.92%.

The improvement was also reflected in hearing-threshold estimation. The number of reports with an exact match to the reference threshold increased from 1169 to 1242, corresponding to an increase in the OK metric from 72.0% to 76.5%. Similarly, the number of reports with threshold estimates within ±10 dB of the reference increased from 1499 to 1530, improving OK_10 from 92.4% to 94.3%. Wave V latency MAE decreased slightly, from 0.29 ms to 0.28 ms.

These findings indicate that the proposed post-processing procedure improves the consistency of intensity-specific detections and contributes to more accurate hearing-threshold estimation. The results also demonstrate that the performance of the complete framework depends not only on the underlying object detector but also on the physiological constraints applied during detection selection.

In the baseline configuration, the highest-confidence detection was selected independently for each intensity level, without applying the cross-intensity physiological constraints.

Table 5. Ablation analysis of the proposed post-processing algorithm using the optimized YOLO26M detector. Changes in percentage-based metrics are expressed in percentage points (pp).

Metric	Without post-processing	With post-processing	Change
TN	2754	2767	+13
TP	2809	2906	+97
FN	375	278	-97
FP	178	165	-13
Accuracy	90.96%	92.76%	+1.80 pp
Precision	94.04%	94.63%	+0.59 pp
Recall	88.22%	91.27%	+3.05 pp
F1-score	91.04%	92.92%	+1.88 pp
Latency MAE	0.29 ms	0.28 ms	-0.01 ms
OK	72.0%	76.5%	+4.5 pp
OK_10	92.4%	94.3%	+1.9 pp

4. Discussion of Results

The results demonstrate that the proposed post-processing algorithm contributes to the performance of the complete ABR interpretation framework. By incorporating physiological constraints across stimulus intensity levels, the algorithm reduced inconsistent detections and improved both wave V identification and hearing-threshold estimation. In particular, the number of false negatives

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

decreased by 25.87%, while the exact hearing-threshold agreement increased from 72.0% to 76.5%, and agreement within ±10 dB increased from 92.4% to 94.3%. These findings indicate that evaluating detections jointly across intensity levels provides an advantage over selecting individual detections without the proposed post-processing constraints.

The comparison of object detection architectures revealed differences in their performance within the complete framework. The optimized YOLO26M model achieved the highest hearing-threshold agreement within ±10 dB (94.3%), followed by the optimized YOLO11M model (94.0%) and the best-performing RF-DETR configuration (93.3%). Although YOLO26M achieved the highest OK_10 score, the differences between the leading configurations were relatively small. Furthermore, the optimized YOLO11M model achieved slightly higher image-level detection accuracy and F1-score than YOLO26M, illustrating that improved performance in individual wave V detections does not necessarily translate directly into better hearing-threshold estimation.

Hyperparameter optimization had a particularly noticeable effect on YOLO26M, increasing its OK_10 score from 92.2% to 94.3%. These findings emphasize the importance of evaluating model configurations using the final clinical task rather than relying exclusively on object detection metrics. Nevertheless, the observed differences should be interpreted as results of the investigated configurations and training procedures rather than as evidence of the general superiority of one detector architecture over another.

Among the RF-DETR configurations, the Large model trained without COCO pretraining achieved a higher OK_10 score than the corresponding pretrained configuration at the same nominal input resolution (93.3% versus 92.9%). This observation suggests that the benefits of natural-image pretraining may depend on the characteristics of the target domain. However, the relatively small difference and the limited number of directly comparable configurations do not allow definitive conclusions regarding the effectiveness of pretraining for ABR graph analysis.

A distinguishing feature of the proposed framework is its direct use of graphical ABR reports rather than reconstructed one-dimensional waveforms. This approach is compatible with clinical workflows in which the available examination records are stored as device-generated PDF reports. The framework simultaneously provides hearing-threshold estimates and intensity-specific wave V latencies, potentially supporting specialists in the interpretation of visually complex recordings. However, the present study did not include a controlled comparison of interpretation time or diagnostic performance between the automated framework and unaided clinicians. Consequently, potential improvements in clinical efficiency and consistency remain to be established through prospective evaluation.

The interpretation of ABR recordings is known to involve inter- and intra-observer variability, particularly when waveform morphology is ambiguous near the hearing threshold. Previous studies have reported inter-rater agreement of approximately 87.3% and intra-rater agreement of approximately 87.9% [37,38]. Although the proposed framework achieved 94.3% agreement with expert-provided hearing thresholds within ±10 dB, these values cannot be compared directly because they represent different evaluation measures and study populations. The present results therefore demonstrate agreement with the available clinical reference annotations, rather than superiority over expert assessment.

An important limitation of the study is the limited heterogeneity of the evaluation data. All recordings originated from a single clinical center, were acquired using one ABR device type, and followed a specific report layout. Although this provided consistent data for model development and evaluation, the generalizability of the framework to other devices, clinical centers, and graphical reporting formats remains unverified. Multi-center and multi-device validation will therefore be necessary before broader clinical deployment can be considered.

A further limitation concerns the reference annotation procedure. Although the ABR reports were annotated by experienced audiologists, detailed information regarding the number of annotators, independent review procedures, disagreement resolution, and inter-rater reliability was unavailable. Consequently, the reference annotations cannot be considered an independently

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

adjudicated consensus standard. Uncertainty in expert interpretation, particularly for responses near the hearing threshold, may have influenced both model development and evaluation. Future studies should incorporate standardized multi-reader annotation protocols and quantify agreement between independent experts.

Taken together, the findings support the feasibility of image-based ABR interpretation and demonstrate the contribution of physiologically constrained post-processing to hearing-threshold estimation. Further studies should address cross-device generalization, reference annotation reliability, and the clinical utility of the framework in routine diagnostic workflows.

5. Conclusions

This study presented a computer-vision-based framework for the simultaneous estimation of hearing thresholds and wave V latencies directly from graphical ABR reports, without reconstructing the underlying one-dimensional signals. The proposed approach combines deep learning object detection with physiologically constrained post-processing to provide a consistent interpretation of responses recorded at different stimulus intensity levels.

The best-performing configuration, based on the optimized YOLO26M model, achieved a wave V detection accuracy of 92.76%, a mean absolute latency error of 0.28 ms, and hearing-threshold agreement with expert-provided reference annotations of 94.3% within a tolerance of ±10 dB. The post-processing algorithm contributed to these results by reducing inconsistent detections and improving hearing-threshold estimation.

These findings demonstrate the feasibility of extracting clinically relevant information directly from ABR report images and support the potential use of the proposed framework as a tool for assisting audiologists in routine interpretation. However, the present evaluation was limited to recordings from a single clinical center and one ABR device type. Further studies involving independent multi-center datasets, different recording systems, and prospective clinical evaluation are needed to establish the generalizability and clinical utility of the approach.

Author Contributions: Writing - Original Draft, Methodology, Validation, Investigation, A.S; Methodology, Software, Validation, Investigation, T.Ś.; Supervision, Writing - review & editing, A.C.; Resources, Data Curation, K.K.; Supervision, H.S.

Funding: Research partially supported by the PolishMedical Research Agency (Institute of Physiology and Pathology of Hearing), grant No. 2024/ABM/03/KPO/KPOD.07.07-IW.07-0222/24-00 ABR.AI.

Institutional Review Board Statement: This work is a part of the broader ABR.AI project. The study was conducted in accordance with the Declaration of Helsinki, and approved by the Bioethics Committee at the Institute of Physiology and Pathology of Hearing (Resolution IFPS:KB.IFPS 6/2025 approved in 18 September 2025).

Informed Consent Statement: The requirement for informed consent was waived by the Bioethics Committee at the Institute of Physiology and Pathology of Hearing.

Data Availability Statement: A public dataset of PNG images with metadata and annotations in CSV format is available in the Kaggle repository (https://www.kaggle.com/datasets/tomaszpiernicki/open-abr) under the CC BY-NC-SA 4.0 license. The original PDF reports are not publicly available due to their sensitive clinical information.

Acknowledgments: The authors are grateful to the staff of the Institute of Physiology and Pathology of Hearing for making available the data analysed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

2D-DWT	Two-dimensional discrete wavelet transform
ABR	Auditory brainstem response
ACC	Accuracy
AE	Linear dichroism
AMLR	Auditory middle latency response
AUC	Area under curve
CNN-LSTM	Convolutional neural network
CWT-ViT	Continuous wavelet transform- Vision Transformer
DCT	Discrete cosine transform
DWT	Discrete wavelet transform
F1	F1-score
FN	False negative
FP	False positive
FPCA	Functional principal component analysis
MAE	Mean absolute error
mAP@0.50	mean Average Precision at IoU = 0.50
mAP@0.50:0.95	mean Average Precision averaged over IoU thresholds 0.50–0.95 (step 0.05)
P	Precision
PDF	Portable document format
R	Recall
SVM	Support vector machine
TN	True negative
TP	True positive

References

1. J. J. Eggermont, 'Chapter 30 - Auditory brainstem response', in Handbook of Clinical Neurology, vol. 160, Elsevier B.V., 2019. doi: https://doi.org/10.1016/B978-0-444-64032-1.00030-8.
2. Y. Liu et al., 'Comparison of Deep Learning Models for Objective Auditory Brainstem Response Detection: A Multicenter Validation Study', Trends in Hearing, vol. 29, p. 23312165251347773, Aug. 2025, doi: 10.1177/23312165251347773.
3. A. Majidpour et al., 'Detection of auditory brainstem response peaks using image processing techniques in infants with normal hearing sensitivity', Biomedical Signal Processing and Control, vol. 86, p. 105117, Sep. 2023, doi: 10.1016/j.bspc.2023.105117.
4. H. Wimalarathna et al., 'Machine learning approaches used to analyze auditory evoked responses from the human auditory brainstem: A systematic review', Computer Methods and Programs in Biomedicine, vol. 226, p. 107118, Nov. 2022, doi: 10.1016/j.cmpb.2022.107118.
5. K. Kochanek, L. Sliwa, J. Zajac, and H. Skarzynski, 'A Universal Computer Audiometer for Objective Hearing Testing and Screening', in 2007 IEEE International Workshop on Medical Measurement and Applications, Warsaw: IEEE, May 2007, pp. 1–3. doi: 10.1109/MEMEA.2007.4285154.
6. A. Erra et al., 'An open-source deep learning-based toolbox for automated auditory brainstem response analyses (ABRA)', Sci Rep, vol. 16, no. 1, p. 9855, Feb. 2026, doi: 10.1038/s41598-026-38045-1.
7. R. M. McKearney, D. M. Simpson, and S. L. Bell, 'Automated wave labelling of the auditory brainstem response using machine learning', International Journal of Audiology, vol. 64, no. 7, pp. 766–771, Jul. 2025, doi: 10.1080/14992027.2024.2404537.
8. R. M. McKearney and R. C. MacKinnon, 'Objective auditory brainstem response classification using machine learning', International Journal of Audiology, vol. 58, no. 4, pp. 224–230, Apr. 2019, doi: 10.1080/14992027.2018.1551633.
9. K. Huang et al., 'Performance Evaluation of Deep Learning Techniques on Classifying Commercial Sensor-based Auditory Brainstem Response', in 2023 16th International Conference on Sensing Technology (ICST), HYDERABAD, India: IEEE, Dec. 2023, pp. 1–4. doi: 10.1109/ICST59744.2023.10460837.
10. F. Özyurt, J. Majidpour, T. A. Rashid, A. Majidpour, and C. Koç, 'Multi-transfer learning techniques for detecting auditory brainstem response', Applied Acoustics, vol. 212, p. 109604, Sep. 2023, doi: 10.1016/j.apacoust.2023.109604.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

11. Y. Liu et al., 'A Deep Learning Model for Automatic Thresholding of Auditory Brainstem Responses: Multicenter and Multispecies Validation', Ear & Hearing, vol. 47, no. 5, pp. 1197–1211, Sep. 2026, doi: 10.1097/AUD.0000000000001811.
12. L. A. Shaheen, B. N. Buran, K. Suthakar, S. D. Koehler, and Y. Chung, 'ABRpresto: An algorithm for automatic thresholding of the Auditory Brainstem Response using resampled cross-correlation across subaverages', Hearing Research, vol. 462, p. 109258, Jun. 2025, doi: 10.1016/j.heares.2025.109258.
13. A. M. Kamerer, 'A Time-Saving Alternative to "Peak-Picking" Algorithms: A Gaussian Mixture Model Feature Extraction Technique for the Neurodiagnostic Auditory Brainstem Response', Ear & Hearing, vol. 45, no. 5, pp. 1115–1124, Sep. 2024, doi: 10.1097/AUD.0000000000001498.
14. N. Acır, Ö. Özdamar, and C. Güzeliş, 'Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection', Engineering Applications of Artificial Intelligence, vol. 19, no. 2, pp. 209–218, Mar. 2006, doi: 10.1016/j.engappai.2005.08.004.
15. Y. Liu et al., 'An interpretable deep learning framework based on time–frequency analysis for automated auditory brainstem response detection', Biomedical Signal Processing and Control, vol. 112, p. 108609, Feb. 2026, doi: 10.1016/j.bspc.2025.108609.
16. Y. Liu et al., 'Derivative-Guided Dual-Attention Mechanisms in Patch Transformer for Efficient Automated Recognition of Auditory Brainstem Response Latency', IEEE Trans. Neural Syst. Rehabil. Eng., vol. 33, pp. 1863–1875, 2025, doi: 10.1109/TNSRE.2025.3558730.
17. S. Liang et al., 'Automatic Recognition of Auditory Brainstem Response Waveforms Using a Deep Learning-Based Framework', Otolaryngol.--head neck surg., vol. 171, no. 4, pp. 1165–1171, Oct. 2024, doi: 10.1002/ohn.840.
18. R. M. McKearney, S. L. Bell, M. A. Chesnaye, and D. M. Simpson, 'Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods', Ear & Hearing, vol. 43, no. 3, pp. 949–960, May 2022, doi: 10.1097/AUD.0000000000001151.
19. F. Ceriani et al., 'A machine-learning-based approach to predict early hallmarks of progressive hearing loss', Hearing Research, vol. 464, p. 109328, Aug. 2025, doi: 10.1016/j.heares.2025.109328.
20. A. Erra et al., 'An Open-Source Deep Learning-Based Toolbox for Automated Auditory Brainstem Response Analyses (ABRA)', Jun. 20, 2024, Neuroscience. doi: 10.1101/2024.06.20.599815.
21. O. Manta, M. Sarafidis, W. Schlee, B. Mazurek, G. K. Matsopoulos, and D. D. Koutsouris, 'Development of Machine-Learning Models for Tinnitus-Related Distress Classification Using Wavelet-Transformed Auditory Evoked Potential Signals and Clinical Data', JCM, vol. 12, no. 11, p. 3843, Jun. 2023, doi: 10.3390/jcm12113843.
22. J. Ma et al., 'Auditory Brainstem Response Data Preprocessing Method for the Automatic Classification of Hearing Loss Patients', Diagnostics, vol. 13, no. 23, p. 3538, Nov. 2023, doi: 10.3390/diagnostics13233538.
23. K. Nowakowski and A. Czyżewski, 'Two-Dimensional Discrete Wavelet Transform for Semi-Automatic Threshold Estimation from Clinical ABR Images', Appl. Sci., vol. 16, no. 18, p. 9264, Sep. 2026, doi: 10.3390/app16189264.
24. A. M. Husain, Illustrated manual of clinical evoked potentials. Springer Publishing Company.
25. J. W. Hall, Handbook of auditory evoked responses. Boston London Toronto: Allyn and Bacon, 1992.
26. C. Chen et al., 'Automatic Recognition of Auditory Brainstem Response Characteristic Waveform Based on Bidirectional Long Short-Term Memory', Front. Med., vol. 7, p. 613708, Jan. 2021, doi: 10.3389/fmed.2020.613708.
27. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, 'End-to-End Object Detection with Transformers', in Computer Vision – ECCV 2020, vol. 12346, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds, in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 213–229. doi: 10.1007/978-3-030-58452-8_13.
28. T.-Y. Lin et al., 'Microsoft COCO: Common Objects in Context', in Computer Vision – ECCV 2014, vol. 8693, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds, in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.
29. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, 'Optuna: A Next-generation Hyperparameter Optimization Framework', in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 30 October 2026 https://doi.org/10.20944/preprints202610.3100.v1

Discovery & Data Mining, Anchorage AK USA: ACM, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

30. R. Khanam and M. Hussain, 'YOLOv11: An Overview of the Key Architectural Enhancements', Oct. 23, 2024, arXiv: arXiv:2410.17725. doi: 10.48550/arXiv.2410.17725.
31. 'Ultralytics YOLO11'. Accessed: Apr. 10, 2026. [Online]. Available: https://docs.ultralytics.com/models/yolo11/
32. 'Ultralytics YOLO26'. Accessed: Apr. 10, 2026. [Online]. Available: https://docs.ultralytics.com/models/yolo26/
33. I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri, 'RF-DETR: Neural Architecture Search for Real-Time Detection Transformers', Nov. 12, 2025, arXiv: arXiv:2511.09554. doi: 10.48550/arXiv.2511.09554.
34. 'RF-DETR: Real-Time SOTA Detection and Segmentation Model'. Accessed: Apr. 10, 2026. [Online]. Available: https://rfdetr.roboflow.com/1.4.0/
35. R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, 'RF-DETR Object Detection vs YOLOv12 : A Study of Transformer-based and CNN-based Architectures for Single-Class and Multi-Class Greenfruit Detection in Complex Orchard Environments Under Label Ambiguity', Apr. 17, 2025, arXiv: arXiv:2504.13099. doi: 10.48550/arXiv.2504.13099.
36. W. He, Y. Zhang, T. Xu, T. An, Y. Liang, and B. Zhang, 'Object Detection for Medical Image Analysis: Insights from the RT-DETR Model', in Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence, Kuala Lumpur Malaysia: ACM, Feb. 2025, pp. 415–420. doi: 10.1145/3730436.3730506.
37. M. Zaitoun, S. Cumming, A. Purcell, and K. O'brien, 'Inter and intra-reader variability in the threshold estimation of auditory brainstem response (ABR) results', Hearing Balance and Communication, vol. 14, no. 1, pp. 59–63, Mar. 2016, doi: 10.3109/21695717.2016.1110957.
38. M. Zaitoun, S. Cumming, and A. Purcell, 'Inter and intra-reader agreement among audiologists in reading auditory brainstem response waves', CANADIAN JOURNAL OF SPEECH-LANGUAGE PATHOLOGY & AUDIOLOGY, vol. Vol. 38, no. No. 4, Winter 2014, [Online]. Available: https://cjslpa.ca/files/2014_CJSLPA_Vol_38/CJSLPA_Winter_2014_Vol_38_No_4.pdf#page=62

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.