IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 18, 2021, accepted July 22, 2021, date of publication August 3, 2021, date of current version August 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102096

On the Use of Machine Learning for Classifying Auditory Brainstem Responses: A Scoping Review

RIDA AL OSMAN[1] AND HUSSEIN AL OSMAN[2], (Member, IEEE)
[1]Faculty of Health Sciences, School of Rehabilitation Sciences, University of Ottawa, Ottawa, ON K1H 8M5, Canada
[2]Faculty of Engineering, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Hussein Al Osman (halosman@uottawa.ca)

ABSTRACT Recent advances in machine learning have led to a surge of interest in classification of the auditory brainstem response. In this work, we conducted a search in the PubMed, Google Scholar, SpringerLink, ScienceDirect, and Scopus databases, and identified twelve studies that explored the use of machine learning to classify the auditory brainstem response as a complementary and objective method to (a) help clinicians better diagnose hearing impairment by discerning between healthy and pathological auditory brainstem response waveforms, (b) present a neural marker for potential applications in hearing aid tuning, and (c) provide a biometric marker for discriminating between subjects. A comparison between the studies presented in this review is not possible as they used different test subjects, group sizes, and stimuli, and evaluated auditory brainstem response differently. Instead, the result of these studies will be presented and their limitations as well as their potential applications will be discussed. Overall, the findings of these studies suggest that ABR classification using machine learning is a promising tool for assessing patients with hearing loss, optimizing technologies for tuning hearing aids, and discriminating between subjects.

INDEX TERMS Auditory brainstem response, classification, decoding, feature extraction, machine learning.

I. INTRODUCTION

Traditionally, audiologists and clinicians used pure tone audiometry to diagnose hearing impairments [1]. They also relied on perceptual assessments such as detection, discrimination, or identification of vowels / consonants in nonsense words and real words to diagnose auditory processing disorders [1]. These assessments are not appropriate for all populations. Infants and children cannot cooperate entirely to disclose their listening experiences, whereas adults with severe hearing loss may have difficulty reporting what they have heard. In addition, poor performance on such behavioral tests could be attributed to other factors such language barrier, wakefulness, mood, and motivation [2]. To overcome these limitations, clinicians and researchers turned to the Auditory Brainstem Response (ABR) which encodes stimulus-specific information with a high degree of accuracy, including timing, the fundamental frequency, and fine structure (harmonics) [2]–[4].

The ABR is a subcortical evoked potential that provides diagnostic information about the pathway from the auditory periphery to the brainstem [5]. It was discovered nearly five decades ago [39], [40] and has been shown to be an effective tool for characterizing the synchronous neural activities in the brainstem [3]. Early ABR studies used stimuli other than speech such as clicks and tones to examine the auditory function; however, such stimuli lacked authenticity considering that they do not replicate sounds from the real world. Therefore, auditory neuroscience has progressively shifted to the use of speech stimuli (e.g. vowels, consonant vowels, and words) [3], [4], [6]–[8]. ABR is typically measured in an anechoic audiometric room using a three electrodes setup where a recording electrode is placed at the vertex, a reference electrode on the right ear lobe, and the ground electrode on the left ear lobe [3]. Speech elicits transient and steady-state neural responses at the level of the brainstem. The transient response refers to the initial portion of the ABR (< 20 ms) which encodes the onset of the neural response to the stimulus [3].

The transient response is commonly assessed in the time domain by analyzing peaks I, II, III, IV, V and A amplitudes and latencies. Peaks V and A (known as the VA complex) are commonly analyzed in clinical settings [3], [9] (see Figure 1). On the other hand, the steady state response

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Tucci.

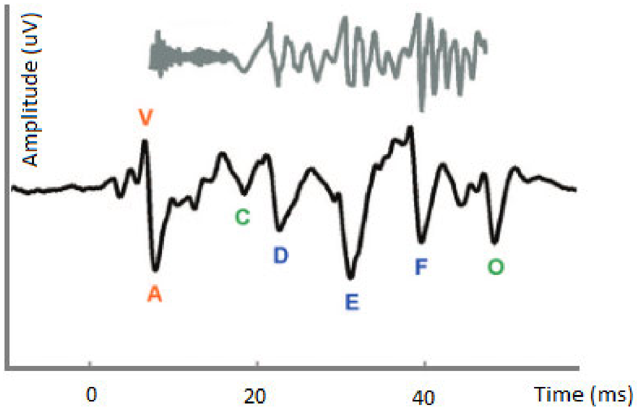

FIGURE 1. Time domain of a 40-ms stimulus /da/ (top) and ABR (bottom). The stimulus evokes characteristic peaks in the transient ABR (V, A, C, and O), and steady-state ABR (D, E, and F). Modified from [3].

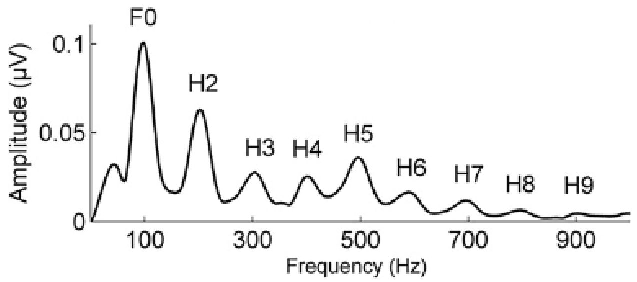

FIGURE 2. Frequency-domain representation of the frequency following response was generated using the fast Fourier transform (FFT) in response to a 40-ms stimulus /da/. Modified from [3].

(> 20ms) is analyzed in the frequency domain using fast Fourier transform (FFT). It is characterized by the envelope frequency response (EFR) and frequency following response (FFR), reflecting the ensemble phase-locked responses to speech stimulus periodicity and its fine structure, respectively [8], [9], [38]. The EFR is analyzed at the fundamental frequency F0 and its harmonics while the FFR is analyzed at the first and/or second formants, up to 1000 Hz as neural phase-locking considerably degrades above this frequency [9]. EFR and FFR diagrams are shown in Figure 2. Several studies have reported that the EFR and FFR correlate with long-term experience with music or tonal language [10], [11], the ability to understand speech in the presence of background noise, [6], [12], [13] and reverberation [14], [16].

As mentioned above, the analysis and interpretation of the transient ABR involves identifying the response waveforms and then measuring their amplitudes and latencies against normative data [3]. This process requires skill and expertise. While clinicians frequently draw similar conclusions; however, differences can occur especially with less experienced audiologists [17].

In addition, the literature does not clearly describe normative data for some hearing impairments, such as auditory processing disorder (APD) [18]. Therefore, there is a need to automate the ABR analysis to support clinicians in making an accurate diagnosis.

Usually, hearing aid fitting is conducted by diagnostic testing using simple stimuli such as clicks or tone pips that do not always achieve optimum results [19], [20]. On the other hand, the use of ABR has proven to achieve better outcomes in selecting hearing aid for patients with hearing impairment. Dajani *et al.* [21] offered insights into how to exploit small changes in speech-evoked ABR waveforms to configure hearing aid parameters such as amplification, and compression levels, for optimal hearing. It is therefore expected that automating the analysis of the ABR waveforms would help patients with hearing aid fitting.

In addition to the clinical benefits, ABR has been recently considered a potential candidate for biometric applications [22]. Currently, face, iris, DNA, and fingerprint recognition are among the most common instruments for biometric subjects' identification. Since individuals have unique ABRs [3], the ABR may also be used for the same purpose.

Artificial intelligence (AI) refers to the use of computers to automate complex tasks generally performed by humans. Machine learning (ML) is a type of AI that makes a prediction or decision by learning patterns in training data and not through direct programming [23].

The use of ML in ABR is a rapidly developing field, predominantly among researchers, neuroscientists, and audiologists, as an objective means to automate the analysis of the speech-evoked neurophysiological data to achieve better diagnosis.

First, the ML classifier is trained by constructing a statistical model with a sample of the ABRs (observations) and their corresponding task classification labels [24]. Subsequently, the remaining sample of the ABRs' data is ingested into the classifier, without labels, and the classifier returns the predicted labels. The performance of the ML classifier is calculated by comparing the labels predicted by the classifier with the ground truth.

To the best of our knowledge, this is the first review that focuses on the recent literature of ABR classification using ML to (1) predict the stimuli labels from the ABRs, (2) automate the analysis of the ABRs, and (3) decode subjects' identity for potential biometric applications.

II. METHODOLOGY

This review followed the Preferred Reporting Elements for Systematic Reviews and Meta-analyses (PRISMA) guidelines [25]. Articles reviewed were originally published in journals indexed in the following databases: PubMed, Google Scholar, SpringerLink, ScienceDirect, and Scopus. The keywords included in the query were: 1. Speech-evoked auditory brainstem response, 2. Click-evoked auditory brainstem response, 3. frequency following response, 4. classification, 5. decoding, and 6. machine learning. The search phrases were created by merging the Medical Subject Headings (MeSH) to locate other relevant machine learning

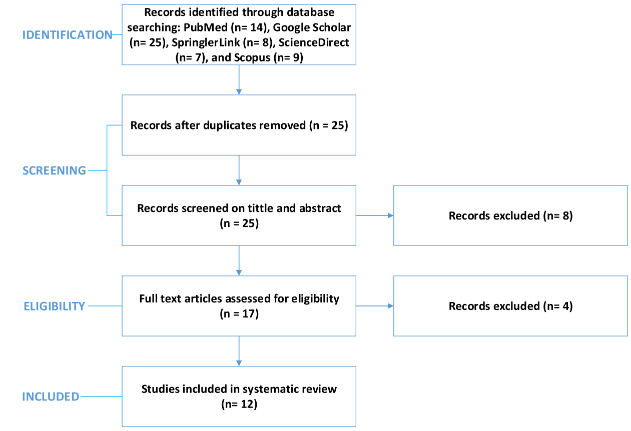
IDENTIFICATION
Records identified through database searching: PubMed (n= 14), Google Scholar (n= 25), SpringlerLink (n= 8), ScienceDirect (n= 7), and Scopus (n= 9)
SCREENING
Records after duplicates removed (n = 25)
Records screened on tittle and abstract (n = 25)
Records excluded (n= 8)
ELIGIBILITY
Full text articles assessed for eligibility (n = 17)
Records excluded (n= 4)
INCLUDED
Studies included in systematic review (n= 12)

FIGURE 3. PRISMA study selection diagram.

keywords. Then truncation was used for the word 'classify*' to look for synonyms or variations on the word stem. Subsequently, the search query was combined with logical operators as follows: (1 OR 2 OR 3) AND (4 OR 5) AND (6).

The criteria used for inclusion included: (i) articles published in the last 5 years (2015-20), (ii) peer-reviewed articles and conference papers, and (iii) articles that focused on classification of ABR. Duplicate studies were eliminated from all three databases. The exclusion criteria consisted of: (i) animal experiments, and (ii) articles that were not published in English.

III. RESULTS

The initial search returned 25 relevant articles. After screening all the eligible papers using the inclusion and exclusion criteria, we ended up with 12 articles (see Figure 1). Among the articles, three used English vowels as a stimuli, three used English consonant-vowels, two used Chinese lexical tones, and the rest of the studies used tone pips, clicks, and musical notes.

Table 1 summarizes the information from the selected articles consisting of: (i) the sample size, (ii) the type of stimulus to evoke the ABR, (iii) the ML classifier, (iv) the ML validation method, (v) the extracted ABR features, (vi) the outcome predicted, (vii) the classifier's performance (accuracy, sensitivity, and specificity rates, and the area under the curve), (viii) the key findings, applications, and limitations.

IV. DISCUSSION

This scoping review examines 12 studies that developed ML models to classify ABR. A comparison among these studies is not possible as they used very different test subjects, group sizes, and stimuli, and evaluated ABR differently. Instead, the results of these studies will be discussed, and their limitations, and the applications they might have, will be highlighted

A. TRADITIONAL ML MODELS TO CLASSIFY ABR

In total, 14 ML classifiers were implemented in all studies presented in this review, with Support Vector Machine (SVM) being the most frequently proposed. Although SVM is regarded as a traditional classifier, it can help reduce overfitting, especially for smaller datasets compared to modern deep learning models. Overfitting occurs when the classifier becomes overly accustomed to the training data to the degree that it has a detrimental effect on its performance with another dataset. Shirzyhiyan *et al.* [26] compared accuracy rates across four traditional ML classifiers (k-nearest Neighbor (KNN), Naive Bayes (NB), Multiclass Support Vector Machine (MSVM) and Discriminant Analysis (DA)) and reported that MSVM achieved the highest accuracy rate – 97.5 % for both combined transient and steady-state features. MSVM may have some advantages over the other classifiers presented in [26]. It is commonly recognized that MSVM can achieve satisfactory performance when the number of features is large and the training dataset size is limited [27], [28]. Other studies used different types of traditional classifiers and their performance scores varied considerably. For example, Llanos *et al.* [22] and Llanos *et al.* [29] used a Hidden Markov Model (HMM) classifier and achieved accuracy rates between 74% and 88% and Area Under the Curve (AUCs) of 0.83 (over different ABR sessions and tones) and 0.93 (same tone and same ABR session), respectively. Yi *et al.* [30] used a Gradient Boosted Decision Tree (GDBT) classifier and obtained an AUC of 0.668. Molina *et al.* [31], Dobrowolski *et al.* [32] used a pattern-based classification and SVM classifiers, respectively, and achieved accuracy, sensitivity and specificity rates of 99.4%, 97.6% and 100%, and 92%, 85%, and 96%, respectively. Xie *et al.* [33] also developed an SVM classifier and obtained accuracy rates between 60-77%. To generalize these results, a larger dataset and a more complex classifier such as Convolutional Neural Network (CNN) is required.

B. CNN-BASED ML MODELS TO CLASSIFY ABR

Only [17], [18] used a CNN classifier. McKeraney and MacKinnon [17] examined 232 paired ABR waveforms of tone pips from eight normal-hearing subjects. The authors classified the paired of ABR waveforms into one of three classes: clear response (CR), absent response (AR), and inconclusive response (IR) based on the decision criteria described in [34]. The authors' findings suggest that CNN could be used in clinical settings, to help audiologists and clinicians interpret ABR waveforms in an objective manner that would allow accurate diagnosis of patients' hearing status. Although this study has its limitations such as small sample size, and use of tone pips instead of speech, and a transient ABR only, the results provide early indications regarding the applicability of such technology in objective clinical hearing assessment.

The performance of CNN classifiers has surpassed that of traditional classifiers for numerous applications. Unlike traditional classifiers, which rely on hand-crafted features that are selected based on the expertise of the developers and/or a tedious trial-error process, CNN classifiers automatically perform feature extraction through the convolution

TABLE 1. **Summary of reviewed studies.**

Study	Sample Size	Stimuli	Classifier	ABR features	Classification Task	Accuracy %	Sensitivity %	Specifity %	Area Under Curve	Key Findings + Applications	Limitations
[20]	8	English vowels (/a/, /o/, /ae/, /i/, and /u/)	LDA	Transient Response (V and A latencies and amplitudes)	/a/ vs /o/ vs /ae/ vs /i/ vs /u/	38.3%				- **Key findings:** This study demonstrated that the speech-evoked ABR of five English vowels (/a/, /o/, /ae/, /i/, and /u/) can be classified using LDA with highest accuracy rate 83.33% obtained with the steady-state responses (EFR + FFR) features. - **Application:** Potential marker to tune hearing aids where evoked responses to confusable speech sounds would be clustered into maximally separated classes.	Small sample size (n= 8); low accuracy rate (38.33 %) for transient response features; the authors did not report the accuracy rate of the combination of the transient and steady-state responses.
				EFR amplitudes (at 100 Hz up to 1000 Hz)		70.8%					
				FFR amplitudes (at 100 Hz up to 1000 Hz)		59.6%					
				EFR+FFR amplitudes (at 100 Hz up to 1000 Hz)		83.3%					
[32]	130	Click	SVM	DWT	Healthy vs pathological	97.0%	95.0%	98.0%		-**Key findings:** This study demonstrated that the click-evoked ABR can be classified using SVM with high accuracy rate of 97%. - **Application:** Potential marker to help audiologists better discriminate between healthy and pathological ABR waveforms	Absence of a speech stimuli.
[31]	83	Click	Pattern Based	ABR time domain series	Healthy vs pathological	97.6%	100%	95.1%		-**Key findings:** This study demonstrated that the click-evoked ABR can be classified using a pattern based classifier with high accuracy rates of 97.6%. - **Application:** Potential marker to help audiologists better discriminate between healthy and pathological ABR waveforms	Absence of a speech stimuli; the proposed method requires expertise in audiology.
[30]	25	English vowels (/æ/, and /u/) from two speakers	GBDT	FFR spectral feature space	/æ/ vs /u/				0.668	**Key findings:** This study demonstrated that a single-trial of the speech-evoked ABR of two English vowels from two speakers can be classified using the Gradient Boosted Decision Tree ML classifier. The authors projected the raw FFRs onto a 12-dimensional spectral feature. Their results showed that across 25 participants, vowel decoding resulted in a mean AUC measure of 0.668 (SD = 0.145). **Applications:** Potential marker to reduce ABR recording sessions	Low accuracy rate.
[37]	22	English vowels (/a/, /ɔ/, /U/, and /u/) at four levels (55, 65, 75, 85 dBA)	SVM	EFR (at F0, H2 to H6) and FFR (at F1, H6, and H8) amplitudes	/a/ vs /ɔ/ vs /U/ vs /u/ at four levels (55, 65, 75, 85 dBA)	80.5%				**Key findings:** This study demonstrated that the speech-evoked ABR of four English vowels (/a/, /ɔ/, /U/, and /u/) presented at four levels (55, 65, 75, 85 dBA) can be classified using SVM. A mean vowel decoding classification accuracy of 80.5% was found at 85 dBA. **Applications:** Potential marker for perceptual discrimination and loudness control for hearing aid users.	Absence of a complex classifier. Accuracy rates for 55, 65, and 75 dBA were not reported.

TABLE 1. *(Continued.)* **Summary of reviewed studies.**

[29]	28	Chinese lexical tones (/yi1/'clothing', /yi2/'aunt' , /yi3/'chair', and /yi4/'easy')	HMM	F0 contours extracted from the FFRs	/yi1/'clothing' vs /yi2/'aunt' vs /yi3/'chair' vs and /yi4/'easy'	74-88%				**Key findings:** This study demonstrated the FFR of four Mandarin lexical tones (/yi1/'clothing', /yi2/'aunt' , /yi3/'chair', and /yi4/'easy') can be classified with accuracy rates ranging from 74% to 88%. **Applications:** Potential marker to assess the robustness of neural pitch encoding, and to complement existing analytical methods that assess auditory function.	Lack of analysis of various frequencies in the FFRs.
[33]	8	Mandarin tones: T1, high-level; T2, low-rising; T4, high-falling	SVM	Spectro-temporal information spanning a narrow frequency band (80–180 Hz) that covers F0 range of all the three tones (100 to 140 Hz)	/T1, high-level/ vs /T2, low-rising/ vs /T4, high-falling/	60-77%				**Key findings:** This study demonstrated that speech-evoked ABR of Mandarin tones (T1, high-level; T2, low-rising; T4, high-falling) recorded in two auditory contexts (predictable or variable) while subjects performed a high and low visual search task can be classified using SVM. The decoding accuracies were higher in the low visual load condition (Mdn = 76.67%, 99th percentile = 83.33%) compared to the high-load condition (Mdn = 60%, 99th percentile = 68.33%) for the predictable auditory context. However, for the variable auditory context, decoding accuracies were significantly lower in the low visual load condition (Mdn = 65%, 99th percentile = 73.33%) relative to the high-load condition (Mdn = 76.67%, 99th percentile = 85%). Their results suggest that additional mechanisms are at play in mediating the impact of visual load on early auditory processing of speech. **Applications:** Provide important insights into the neural mechanisms of multisensory processing and suggest that crossmodal influences can potentially be recognized during the brainstem encoding stage.	Small sample size.
[26]	27	Consonant-vowel syllables (/ba/, /da/, and /ga/)	KNN	Time domain Features	/ba/ vs /da/ vs /ga/	45.0%	43.4%			**Key findings:** This study demonstrated that the speech-evoked ABR of three consonant-vowel syllables (/ba/, /da/, and /ga/) can be classified using four different ML classifiers (DA, MSVM, KNN, and NB). The authors achieved highest	Time domain response features did not achieve high accuracy rate (43 -50 % across all four classifiers).
			NB			44.7%	44.8%				
			DA			48.4%	43.8%				
			MSVM			50.3%	48.4%				
			KNN	Amplitude Frequency		58.0%	56.7%				
			NB			58.3%	56.9%				

TABLE 1. *(Continued.)* **Summary of reviewed studies.**

			DA	domain features		58.9%	57.7%			accuracy and sensitivity rates of 97.5 %, 97.7 %, respectively using MSVM when combined features from time and frequency. domains were added together. **Applications:** Potential marker to reduce ABR recording sessions.	
			MSVM			61.1%	59.6%				
			KNN	Phase frequency domain features		80.2%	80.2%				
			NB			70.9%	70.8%				
			DA			74.6%	74.8%				
			MSVM			77.4%	77.4%				
			KNN	Time-frequency domain features		91.3%	93.1%				
			NB			71.6%	73.7%				
			DA			72.5%	73.7%				
			MSVM			89.5%	89.6%				
			KNN	Combined features from time, frequency, and time-frequency domains.		96.3%	96.3%				
			NB			84.8%	85.0%				
			DA			88.8%	88.8%				
			MSVM			97.5%	97.7%				
[17]	8	1 kHz and 4 kHz tone pips with a 2-cycle rise/fall time	CNN	Wave V of the transient response	/clear response/ vs /inconclusive response/ vs /response absent/	92.9%	92.9%	96.4%		**Key findings:** This study demonstrated that tone pips evoked ABR can be classified using CNN. The authors provided evidence that a CNN classifier is able to learn from the decision making process of clinicians to classify ABR waveforms to an accuracy of 92.9%. **Applications**: Potential marker to help clinicians with less experience in ABR waveform analysis, and provide an objective diagnosis for newborn hearing screening.	Small sample size; absence of speech stimuli; wave A of the transient response was not analyzed.
[24]	13	(/ba/, /da/, /di/, and three musical notes)	LDA	Time domain	/ba/ vs /da/ vs /di/ vs musical notes	62.8%				**Key findings:** This study demonstrated that speech-evoked ABR of three CV phones (/ba/, /da/, and /ga/) and three musical notes can be classified using LDA. The authors reported accuracy rates of 62.8%, 61.8%, 39.7%, and 71.5% in time domain, frequency domain amplitude, frequency domain phase, and frequency domain combined amplitude and phase. **Applications:** Potential marker to better diagnose patients with hearing loss.	Small sample size; moderate accuracy rate.
				Frequency domain Amplitudes (up to 1000 Hz)		61.8%					
				Frequency domain phase angle (up to 1000 Hz)		39.7%					
				Combined frequency domain amplitude and phase		71.5%					
[22]	20	Mandarin tones (/T1/, high-level; /T2/, low-rising; /T4/, high-falling)	HMM	Same auditory context (i.e., same tone and session)	/excellent (AUROC > 0.9)/ vs /good (> 0.8)/ vs /fair (> 0.7)/ vs /poor (> 0.6)/ vs /bad (> 0.5)/				0.93	**Key findings**: This study demonstrated that speech-evoked ABR can be classified using HMM to discriminate between subjects. Two group of subjects were included in the study: Native speakers of English and Mandarin. The authors decoded subjects' identity within the same auditory context (same tone and session) and across different stimuli and recording sessions with AUROC of 0.93, and 0.83, respectively. **Applications**: Potential marker for biometric subjects identification	The authentication of subjects across different days could be improved if the authors were to train a full-bandwidth model (1 – 1000 Hz) with FFRs from different days. The authors reported that a quick simulation of this setting showed an improvement from 0.77 (trained with the first session, tested with the third session) to 0.82 (trained with the first two sessions, tested with the third session). Another limitation is stimulus familiarization which could have a negative impact on the classifier performance.
				Across different sessions and tones (i.e., across different auditory contexts)					0.83		

TABLE 1. ***(Continued.)*** **Summary of reviewed studies.**

[18]	136	click	SVM	Features in the time-frequency representation	Healthy vs pathological	91.0%				**Key findings:** This study demonstrated that click-evoked ABR can be classified with SVM, RF, DT, GB, Xgboost, and NN to discriminate between healthy and pathological ABR waveforms. The authors reported that the Xgboost classifier returned highest accuracy rate (92%) among all classifiers. **Applications:** Potential marker for the future development of an automatic evaluation tool for clinical ABR waveform analysis.	Absence of speech stimuli.
			RF			90.0%					
			DT			84.0%					
			GB			86.0%					
			Xgboost			92.0%					
			NN			84.0%					

Abbreviations:
DWT (Discrete Wavelet Transform); ABR (Auditory Brainstem Response); EFR (Envelope Frequency Response); FFR (Frequency Following Response); LDA (Linear Discriminant Analysis); DA (Discriminant Analysis); (SVM (Support Vector Machine); MSVM (Multiclass Support Vector Machine); GDBT (Gradient Boosted Decision Tree); KNN (k-nearest Neighbor); NB (Naïve Bayes); HMM (Hidden Markov Model); CNN (Convolutional Neural Network), NN (Neural Network), RF (Random Forest); GB (Gradient Boosted); DT (Decision Tree); Xgboost (Extreme Gradient Boosting).

of the input information with learned filters at one or more hidden layers. By increasing the number of convolutional layers, the network can learn to extract complex features at its deepest layers. It is due to these learned filters that CNN architectures tend to perform better than traditional classifiers. However, given the number of learned parameters throughout the network, CNN classifiers require vastly larger datasets to achieve their full potential. In the case of ABR records, CNN can be directly fed the pre-processed data without further feature extraction. Hence, the focus for developing a CNN-based model would be at selecting the most appropriate architecture rather than the optimal features.

C. HOW MUCH DATA IS ENOUGH?

One critical question remains about the use of ML to classify ABR data: How much data is enough? Although a definitive answer cannot be determined, some ML models are more permissive or tolerable for small samples. For example, both LDA and SVM can offer satisfactory performance for small training datasets and are less likely to overfit while CNN requires a large sample size. One method for predicting sample size is to use cortical EEG ML based classification studies as a reference. In this context, Llanos *et al.* [22] examined several studies that classified cortical EEG with ML models and found that an average sample size of 20 subjects is sufficient for traditional ML algorithms. In our review, we apply this estimation as a criteria; all studies that measured less than 20 subjects were considered to have a small sample size as we report in Table 1 (under the Limitations column).Some of the studies presented in this review indicated that additional data might improve their classification algorithms. However, considering the limitations of subject recruitment and long ABR recording sessions, increasing the sample size is not always feasible. These limitations can be mitigated by making research data publicly available so others can use or combine it with their existing data.

D. HOW TO ENSURE GOOD QUALITY OF THE DATA?

It is not only the quantity of data that is important for high-performance ML classification, but also the quality of the data is critical. Because ABR data can easily accumulate noise at multiple stages during collection, ML prediction could be highly impacted especially when CNN classifiers are used [43]. As a result, it is critical to minimize noise during the ABR collection process. Al Osman *et al.* [16] discuss noise reduction procedures in detail, as well as how to prepare participants for optimal ABR data collection. To summarize, both the quantity and quality of ABR data are critical factors for best ML classification performance.

E. APPLICATIONS OF ML MODELS TO CLASSIFY ABR

Eight of the 12 studies presented in this review suggested that classification of the ABR can be used to diagnose hearing impairments and auditory processing disorders, perform biometric identification of subjects, or to adjust hearing aid configuration. However, they did not elaborate on the feasibility of such applications. Only four of the studies detailed the potential application of their technology. For example, Llanos *et al.* [22] demonstrated that classification of FFRs could be used for biometric applications in the context of subject-discrimination tasks, especially considering their simple three electrodes setup. This is a novel finding as previous EEG biometric identification studies were conducted exclusively with cortical EEGs.

Furthermore, Molina *et al.* [31] developed an ML classifier to decode the ABR time series of 83 young subjects based on symbolic pattern discovery. Their findings offer insights into potential applications to predict whether patients have an auditory-related disorder. Dobrowolski *et al.* [32]

compared six ML classifiers (SVM, RF, DT, GB, EGB, and NN) and found that EGB was the most robust algorithm with an accuracy value of 92%. Their results suggest that EGB may be well suited for the future development of an automatic evaluation tool for clinical ABR waveform analysis. Taken together, the findings of Wimalarath *et al.* [18] and Dobrowolski *et al.* [32] suggest that automating speech-evoked ABR analysis using machine learning may provide a complementary means to help clinicians better diagnose their patients' hearing. On the other hand, a significant limitation for these studies is the size of the dataset available. With sufficient data, ML models could perform better than humans in a variety of complex tasks [39], [40]. McKeraney and MacKinnon [17] suggested that if a machine learning model was trained on sufficient data labelled by a group of experts, the model's predictions would improve and potentially match those of the expert group. Such algorithm could then be embedded as a module in the ABR software to provide clinicians with real-time assistance in ABR analysis.

F. QUALITY ASSESSMENT

To assess the quality of the twelve targeted studies, the reporting items from the checklist described in Table 1 from [44] were compared to the content of the publications. This checklist requested information about the study's nature, objectives, rationale, data collection setup, machine learning models and algorithms, theoretical claims, datasets, validation metrics, experimental results, clinical applications, limitations, and unexpected results. We calculated a mean score of 10.25 ± 1.22, and score disagreements were resolved through consensus.

G. LIMITATIONS

Considering that the twelve studies used vastly different test subjects, group sizes, and stimuli, and evaluated ABR differently, a comparison of their findings was not possible. Instead, a summary of the findings from these studies, their limitations, and potential applications is shown in Table 1.

H. FUTURE WORK

First, future studies should have a larger sample size and take advantage of modern deep learning classifiers such as CNN or combine CNN with traditional classifiers in an ensemble or to create a cascade architecture. Most of the studies included in this review reported difficulties in attaining a large sample size. This constraint could be alleviated by making ABR data public so that others can use it and / or combine it with additional data. Moreover, future studies should employ natural speech stimuli simulated in noise and reverberation to ensure that we are faithfully replicating real-world listening conditions. Furthermore, future research should examine the possibility of using ML to establish a correlation between ABR and auditory thresholds in a variety of populations, including children and the elderly with and without hearing loss, as well as those with and without cochlear implants. Most of the studies covered in this review included young adult participants with normal hearing. To generalize the effectiveness of ML techniques, different populations with various types of hearing impairment are needed to ensure good clinical applicability and effectiveness.

V. CONCLUSION

This review presented 12 studies that explored the use of ML models to classify ABR as a complementary and objective method to (1) support clinicians to better diagnose hearing impairment by discriminating between healthy and pathological ABR waveforms, (2) provide a neural marker to optimize technologies used in hearing aids and cochlear implants, and (3) provide a biometric marker for discriminating between subjects. However, certain challenges must be addressed before these can be established.

REFERENCES

[1] American Speech-Language-Hearing Association. (2005). *(Central) Auditory Processing Disorders: The Role of the Audiologist*. [Online]. Available: http://www.asha.org/members/deskref-journals/

[2] J. Hornickel, B. Chandrasekaran, S. Zecker, and N. Kraus, "Auditory brainstem measures predict reading and speech-in-noise perception in school-aged children," *Behav. Brain Res.*, vol. 216, no. 2, pp. 597–605, Jan. 2011, doi: 10.1016/j.bbr.2010.08.051.

[3] E. Skoe and N. Kraus, "Auditory brain stem response to complex sounds: A tutorial," *Ear Hearing*, vol. 31, no. 3, pp. 302–324, Jun. 2010, doi: 10.1097/aud.0b013e3181cdb272.

[4] A. Koravand, R. A. Osman, V. Rivest, and C. Poulin, "Speech-evoked auditory brainstem responses in children with hearing loss," *Int. J. Pediatric Otorhinolaryngol.*, vol. 99, pp. 24–29, Aug. 2017, doi: 10.1016/j.ijporl.2017.05.010.

[5] B. Chandrasekaran and N. Kraus, "The scalp-recorded brainstem response to speech: Neural origins and plasticity," *Psychophysiology*, vol. 47, no. 2, pp. 236–246, Mar. 2010, doi: 10.1111/j.1469-8986.2009.00928.x.

[6] R. Al Osman, C. Giguère, and H. Dajani, "Effects of stimulus rate and noise on speech-evoked auditory brainstem responses," *Can. J. Speech-Lang. Pathol. Audiol.*, vol. 40, no. 2, pp. 1–14, 2016.

[7] R. A. Osman, "A study of auditory speech processing using brainstem evoked responses under the effects of stressors," Ph.D. dissertation, School Rehabil. Sci., Univ Ottawa, Ottawa, ON, Canada, 2016.

[8] R. Al Osman and H. Al Osman, "Inter-modality influence on the brainstem using an arithmetic exercise," *J. Acoust. Soc. Amer.*, vol. 144, no. 1, pp. EL26–EL32, Jul. 2018, doi: 10.1121/1.5045191.

[9] S. Aiken and T. Picton, "Envelope and spectral frequency-following responses to vowel sounds," *Hear Res.*, vol. 245, nos. 1–2, pp. 35–47, 2008, doi: 10.1016/j.heares.2008.08.004.

[10] A. Krishnan, J. T. Gandour, G. M. Bidelman, and J. Swaminathan, "Experience-dependent neural representation of dynamic pitch in the brainstem," *NeuroReport*, vol. 20, no. 4, pp. 408–413, Mar. 2009, doi: 10.1097/wnr.0b013e3283263000.

[11] G. M. Bidelman, J. T. Gandour, and A. Krishnan, "Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch," *Brain Cogn.*, vol. 77, no. 1, pp. 1–10, Oct. 2011, doi: 10.1016/j.bandc.2011.07.006.

[12] S. Anderson, A. Parbery-Clark, T. White-Schwoch, and N. Kraus, "Development of subcortical speech representation in human infants," *J. Acoust. Soc. Amer.*, vol. 137, no. 6, pp. 3346–3355, Jun. 2015, doi: 10.1121/1.4921032.

[13] G. Musacchia, S. Ortiz-Mantilla, C. P. Roesler, S. Rajendran, J. Morgan-Byrne, and A. A. Benasich, "Effects of noise and age on the infant brainstem response to speech," *Clin. Neurophysiol.*, vol. 129, no. 12, pp. 2623–2634, Dec. 2018, doi: 10.1016/j.clinph.2018.08.005.

[14] G. M. Bidelman and A. Krishnan, "Effects of reverberation on brainstem representation of speech in musicians and non-musicians," *Brain Res.*, vol. 1355, pp. 112–125, Oct. 2010, doi: 10.1016/j.brainres.2010.07.100.

[15] R. Al Osman, H. R. Dajani, and C. Giguère, "Self-masking and overlap-masking from reverberation using the speech-evoked auditory brainstem response," *J. Acoust. Soc. Amer.*, vol. 142, no. 6, pp. EL555–EL560, Dec. 2017, doi: 10.1121/1.5017522.

[16] R. A. Osman, C. Giguère, and H. R. Dajani, "Effects of early- and late-arriving room reflections on the speech-evoked auditory brainstem response," *J. Amer. Acad. Audiol.*, vol. 29, no. 2, pp. 95–105, Feb. 2018, doi: 10.3766/jaaa.16017.

[17] R. M. McKearney and R. C. MacKinnon, "Objective auditory brainstem response classification using machine learning," *Int. J. Audiol.*, vol. 58, no. 4, pp. 224–230, Apr. 2019, doi: 10.1080/14992027.2018.1551633.

[18] H. Wimalarathna, S. Ankmnal-Veeranna, C. Allan, S. K. Agrawal, P. Allen, J. Samarabandu, and H. M. Ladak, "Comparison of machine learning models to classify auditory brainstem responses recorded from children with auditory processing disorder," *Comput. Methods Programs Biomed.*, vol. 200, Mar. 2021, Art. no. 105942, doi: 10.1016/j.cmpb.2021.105942.

[19] K. L. Johnson, T. G. Nicol, and N. Kraus, "Brain stem response to speech: A biological marker of auditory processing," *Ear Hearing*, vol. 26, no. 5, pp. 424–434, Oct. 2005, doi: 10.1097/01.aud.0000179687.71662.6e.

[20] A. Sadeghian, H. R. Dajani, and A. D. C. Chan, "Classification of speech-evoked brainstem responses to English vowels," *Speech Commun.*, vol. 68, pp. 69–84, Apr. 2015, doi: 10.1016/j.specom.2015.01.003.

[21] H. R. Dajani, B. P. Heffernan, and C. Giguère, "Improving hearing aid fitting using the speech-evoked auditory brainstem response," presented at the 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., Osaka, Japan, Jul. 2013.

[22] F. Llanos, Z. Xie, and B. Chandrasekaran, "Biometric identification of listener identity from frequency following responses to speech," *J. Neural Eng.*, vol. 16, no. 5, Jul. 2019, Art. no. 056004, doi: 10.1088/1741-2552/ab1e01.

[23] J. Senders, P. C. Staples, A. V. Karhade, M. M. Zaki, W. B. Gormley, M. L. D. Broekman, T. R. Smith, and O. Arnaout, "Machine learning and neurosurgical outcome prediction: A systematic review," *World Neurosurg.*, vol. 109, pp. 476–486, Jan. 2018, doi: 10.1016/j.wneu.2017.09.149.

[24] S. Losorelli, B. Kaneshiro, G. A. Musacchia, N. H. Blevins, and M. B. Fitzgerald, "Factors influencing classification of frequency following responses to speech and music stimuli," *Hearing Res.*, vol. 398, Dec. 2020, Art. no. 108101, doi: 10.1016/j.heares.2020.108101.

[25] J. Eden, L. Lebitt, A. Berg, and S. Morton, *Finding What Works in Health Care: Standards for Systematic Reviews*. 2011, doi: 10.17226/13059.

[26] Z. Shirzhiyan, E. Shamsi, A. S. Jafarpisheh, and A. H. Jafari, "Objective classification of auditory brainstem responses to consonant-vowel syllables using local discriminant bases," *Speech Commun.*, vol. 114, pp. 36–48, Nov. 2019, doi: 10.1016/j.specom.2019.09.003.

[27] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2007, doi: 10.1007/s10462-007-9052-3.

[28] C. Lehmann, T. Koenig, V. Jelic, L. Prichep, R. E. John, L.-O. Wahlund, Y. Dodge, and T. Dierks, "Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG)," *J. Neurosci. Methods.*, vol. 161, no. 2, pp. 342–350, 2007, doi: 10.1016/j.jneumeth.2006.10.023.

[29] F. Llanos, Z. Xie, and B. Chandrasekaran, "Hidden Markov modeling of frequency-following responses to Mandarin lexical tones," *J. Neurosci. Methods*, vol. 291, pp. 101–112, Nov. 2017, doi: 10.1016/j.jneumeth.2017.08.010.

[30] H. G. Yi, Z. Xie, R. Reetzke, A. G. Dimakis, and B. Chandrasekaran, "Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach," *Brain Behav.*, vol. 7, no. 6, Jun. 2017, Art. no. e00665, doi: 10.1002/brb3.665.

[31] M. E. Molina, A. Perez, and J. P. Valente, "Classification of auditory brainstem responses through symbolic pattern discovery," *Artif. Intell. Med.*, vol. 70, pp. 12–30, Jun. 2016, doi: 10.1016/j.artmed.2016.05.001.

[32] A. Dobrowolski, M. Suchocki, K. Tomczykiewicz, and E. Majda-Zdancewicz, "Classification of auditory brainstem response using wavelet decomposition and SVM network," *Biocybern. Biomed. Eng.*, vol. 36, no. 2, pp. 427–436, 2016, doi: 10.1016/j.bbe.2016.01.003.

[33] Z. Xie, R. Reetzke, and B. Chandrasekaran, "Taking attention away from the auditory modality: Context-dependent effects on early sensory encoding of speech," *Neuroscience*, vol. 384, pp. 64–75, Aug. 2018, doi: 10.1016/j.neuroscience.2018.05.023.

[34] G. J. Sutton and G. Lightfoot. (Mar. 2013). *Guidance for Auditory Brainstem Response Testing in Babies, Version 2.1*. [Online] Available: https://www.thebsa.org.uk/wpcontent/uploads/2014/08/NHSP_ABRneonate_2014.pdf

[35] J. V. Carter, J. Pan, S. N. Rai, and S. Galandiuk, "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves," *Surgery*, vol. 159, no. 6, pp. 1638–1645, Jun. 2016, doi: 10.1016/j.surg.2015.12.029.

[36] S. Thongsuwan, S. Jaiyen, A. Padcharoen, and P. Agarwal, "ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost," *Nucl. Eng. Technol.*, vol. 53, no. 2, pp. 522–531, Feb. 2021, doi: 10.1016/j.net.2020.04.008.

[37] B. Heffernan, H. Dajani, and C. Gigurere, "Towards developing a brain-computer interface for automatic hearing aid fitting based on the speech-evoked frequency following response," presented at the Neurotechnix, Funchal, Portugal, 2017.

[38] J. Krizman and N. Kraus, "Analyzing the FFR: A tutorial for decoding the richness of auditory function," *Hearing Res.*, vol. 382, Oct. 2019, Art. no. 107779, doi: 10.1016/j.heares.2019.107779.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[40] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, and L. Uhlmann, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018, doi: 10.1093/annonc/mdy166.

[41] D. L. Jewett and J. S. Williston, "Auditory-evoked far fields averaged from the scalp of humans," *Brain*, vol. 94, no. 4, pp. 681–696, 1971, doi: 10.1093/brain/94.4.681.

[42] G. Moushegian, A. L. Rupert, and R. Stillman, "Scalp-recorded early responses in man to frequencies in the speech range," *Electroencephalogr. Clin. Neurophysiol.*, vol. 35, no. 6, pp. 665–667, 1973, doi: 10.1016/0013-4694(73)90223-x.

[43] S. Gupta and A. Gupta, "Dealing with noise problem in machine learning data-sets: A systematic review," *Proc. Comput. Sci.*, vol. 161, pp. 466–474, Jan. 2019, doi: 10.1016/j.procs.2019.11.146.

[44] A. M. Maitín, A. J. García-Tejedor, and J. P. R. Muñoz, "Machine learning approaches for detecting Parkinson's disease from EEG analysis: A systematic review," *Appl. Sci.*, vol. 10, no. 23, p. 8662, Dec. 2020, doi: 10.3390/app10238662.

RIDA AL OSMAN received the bachelor's degree in electrical engineering, the master's degree in systems science, and the Ph.D. degree in rehabilitation science (neuroscience) from the University of Ottawa. He was an Adjunct Professor at the School of Rehabilitation Science, University of Ottawa, between July 2017 and July 2020. He is also an author of several conference papers and journal articles.

HUSSEIN AL OSMAN (Member, IEEE) received the bachelor's, master's, and Ph.D. degrees from the University of Ottawa. He is currently an Associate Professor at the University of Ottawa and leads the Multimedia Processing and Interaction Group. He has recently been interested in remote physiological signal measurement and its application in affective computing. He has produced over 50 peer-reviewed research articles, two patents, and several technology transfers to the industry. His research interests include affective computing, specifically multimodal affect estimation, human–computer interaction, serious gaming, and multimedia systems.